(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0023859
(43) 공개일자 2025년02월18일

(51) 국제특허분류(Int. Cl.)
*G06V 10/80* (2022.01) *G06N 3/08* (2023.01)
*G06T 7/73* (2017.01) *G06V 10/40* (2022.01)
*G10K 11/18* (2023.01)
(52) CPC특허분류
*G06V 10/806* (2023.08)
*G06N 3/08* (2023.01)
(21) 출원번호 10-2023-0105118
(22) 출원일자 2023년08월10일
심사청구일자 2023년08월10일

(71) 출원인
**경희대학교 산학협력단**
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
**김정욱**
경기도 안양시 동안구 귀인로 258, 106동 802호 (평촌동,꿈마을라이프아파트)
**김동진**
경기도 수원시 영통구 영일로6번길 47, 405호 (영통동)
**엄성진**
경기도 수원시 팔달구 권광로196번길 53-9, 601호 (인계동)
(74) 대리인
**두호특허법인**

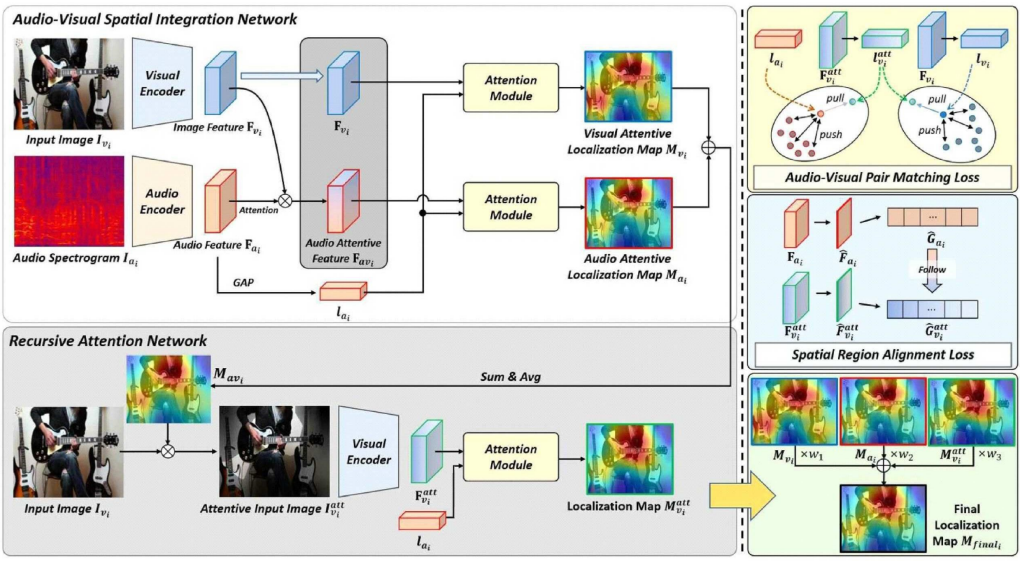
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **시청각의 공간 정보를 이용한 음원 위치 추정 방법 및 이를 수행하기 위한 컴퓨팅 장치, 컴퓨터 프로그램**

## (57) 요 약

본 발명은 인공 신경망 기반의 객체 검출 기술로서, 청각 정보 기반의 객체 검출 기술에 관한 것으로, 이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 단계, 상기 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 단계, 상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 단계, 상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 단계 및 생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 예측하는 단계를 포함한다.

## 대 표 도



(52) CPC특허분류
*G06T 7/73* (2017.01)
*G06V 10/40* (2023.08)
*G10K 11/18* (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711197142
과제번호 00252391
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 멀티모달 데이터가 온전히 주어지지 않는 자율주행환경에서 강인한 객체 검출을 위한 인간의 뇌를 모방한 메모리 구조에 관한 연구
기 여 율 1/1
과제수행기관명 경희대학교산학협력단
연구기간 2023.06.01 ~ 2024.02.29

# 명 세 서

## 청구범위

### 청구항 1

하나의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서,

이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 단계;

상기 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 단계;

상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 단계;

상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 단계; 및

생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 예측하는 단계;를 포함하는, 음원 위치 추정 방법.

### 청구항 2

제 1 항에 있어서,

집중 모듈 통과시 생성된 손실함수, 시청각 맞춤 손실 함수, 공간 영역 조정 손실함수를 합산하여 최종 학습 손실함수를 도출하는 단계; 를 더 포함하는, 음원 위치 추정 방법.

### 청구항 3

제 2 항에 있어서,

상기 최종 학습 손실함수를 적용하는 단계

시각 특징점과 청각 특징점이 집중된 시각 특징점을 따라가도록 손실 함수를 적용하는, 음원 위치 추정 방법.

### 청구항 4

제 3 항에 있어서,

상기 손실 함수는, 하기 수학식으로 나타내는, 음원 위치 추정방법.

(수학식)

$$D(\alpha,\beta)=||(\alpha-\beta)/\tau||_2^2$$

$$T(l_{v_j}^{att},l_{a_i},l_{a_j})=D(l_{vi}^{att},l_{a_i})+\max(\delta-D(l_{vi}^{att},l_{a_i}),0)$$

$$T(l_{v_j}^{att},l_{v_i},l_{v_j})=D(l_{vi}^{att},l_{v_i})+\max(\delta-D(l_{vi}^{att},l_{v_i}),0)$$

δ, τ: 값 조정을 위한 초매개변수(hyper parameter)

$I_{v_i}^{att}$ : 재귀적 시각 특징점의 GAP(Global Average Pooling) 값

$I_v$: 시각 특징점의 GAP(Global Average Pooling) 값

$I_a$: 청각 특징점의 GAP(Global Average Pooling) 값

D(α,β) : α와 β사이의 유클리드 거리(Euclidean distance)

**청구항 5**

제 3 항에 있어서,

상기 최종 학습 손실함수를 적용하는 단계는,

상기 손실 함수를 바탕으로 하기 수학식에 의해 시청각 맞춤 손실 함수(Audio-Visual Pair Matching Loss)를 도출하는, 음원 위치 추정방법.

(수학식)

$$L_{avpm}=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1(j\neq i)}^{N}T(I_{v_i}^{att},I_{a_i},I_{a_j})+T(I_{v_i}^{att},I_{v_i},I_{v_j})$$

N : 학습시 입력되는 영상 데이터 수

**청구항 6**

제 2 항에 있어서,

상기 최종 학습 손실 함수를 적용하는 단계는,

하기 수학식에 의해 청각 특징점이 공간적 정보를 가지도록 공간 영역 조정 손실 함수를 적용하는, 음원 위치 추정 방법.

(수학식)

$$L_{sra}=\frac{1}{N}\sum_{i=1}^{N}D_{KL}(\hat{G}_{v_i}^{att}\parallel\hat{G}_{a_i})$$

N : 학습시 입력되는 영상 데이터 수

$\hat{G}_v^{att}$ : 재귀적 시각 특징점을 채널 요소 평균(Channel-wise Average Pooling)을 한 후 1차원 벡터화(Flatten)

$\hat{G}_a$: 청각 특징점을 채널 요소 평균(Channel-wise Average Pooling)을 한 후 1차원 벡터화(Flatten)

**청구항 7**

제 1 항에 있어서,

상기 재귀적 위치 지역화 맵을 생성하는 단계는,

상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 시청각 위치 지역화 맵을 생성하는 단계,

생성된 시청각 위치 지역화 맵을 입력받은 이미지 데이터와 산술 연산(element-wise)수행하여 집중된 이미지를 생성하는 단계, 및

상기 생성된 집중된 이미지를 시각 딥러닝 네트워크 및 집중 모듈에 통과시켜 재귀적 위치 지역화 맵을 생성하는 단계를 포함하는, 음원 위치 추정 방법.

**청구항 8**

제 7 항에 있어서,

상기 집중된 이미지를 생성하는 단계는,

산술 연산 결과로 입력 이미지 중 음원 출력이 포함된 객체가 있을 것으로 파악되는 영역을 관심 영역으로 생성하는, 음원 위치 추정 방법.

**청구항 9**

제7 항에 있어서,

상기 최종 음원 위치 지역화 맵을 예측하는 단계는,

생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 각각의 가중치를 고려하여 융합하는, 음원 위치 추정 방법.

**청구항 10**

제 1 항에 있어서,

상기 시각 위치 지역화 맵을 생성하는 단계는,

상기 획득된 영상 데이터에서 추출된 이미지 데이터를 시각 딥러닝 네트워크에 통과시켜 시각 특징점을 도출하는, 음원 위치 추정 방법.

**청구항 11**

제 10 항에 있어서,

상기 청각 위치 지역화 맵을 생성하는 단계는,

상기 획득된 영상 데이터에서 추출된 오디오 데이터를 청각 딥러닝 네트워크에 통과시켜 청각 특징점을 도출하는, 음원 위치 추정 방법.

**청구항 12**

제 11 항에 있어서,

상기 청각 위치 지역화 맵을 생성하는 단계는,

상기 도출된 청각 특징점을 상기 도출된 시각 특징점에 곱하여 오디오 중점 특징점(Audio Attentive Feature)을 생성하고, 생성된 오디오 중점 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵(Audio Attentive Localization Map)을 생성하는, 음원 위치 추정 방법.

**청구항 13**

제 1 항에 있어서,

상기 영상 데이터를 획득하는 단계는,

상기 획득된 영상 데이터에서 이미지 촬영 시점을 기준으로 소정 시간 전후의 오디오 데이터를 획득하는, 음원 위치 추정 방법.

**청구항 14**

제 13 항에 있어서,

상기 영상 데이터를 획득하는 단계는,

상기 획득된 오디오 데이터를 변환한 로그-스케일 스텍트로그램 정보를 생성하는, 음원 위치 추정 방법.

**청구항 15**

하나 이상의 프로세서들, 및

메모리; 및

하나 이상의 프로그램들을 포함하고,

상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며,

이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 영상 획득 모듈;

상기 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 시각 위치 지역화 맵 생성 모듈;

상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 청각 위치 지역화 맵 생성 모듈;

상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 재귀적 위치 지역화 맵 생성 모듈; 및

생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 생성하는 최종 음원 위치 지역화 맵 생성 모듈을 포함하는, 컴퓨팅 장치.

**청구항 16**

제 15 항에 있어서,

집중 모듈 통과시 생성된 손실함수, 시청각 맞춤 손실 함수, 공간 영역 조정 손실함수를 합산하여 최종 학습 손실함수를 도출하는 손실함수 도출 모듈;을 더 포함하는, 컴퓨팅 장치.

**청구항 17**

제 16 항에 있어서,

상기 손실함수 도출 모듈은,

청각 특징점과 시각 특징점이 집중된 시각 특징점을 닮아가도록 하는 시청각 맞춤 손실 함수 도출 모듈 및

청각 특징점이 공간적인 정보를 가질 수 있도록, 집중된 시각 특징점을 닮아가도록 하는 공간 영역 조정 손실함수 도출 모듈을 포함하는, 컴퓨팅 장치.

**청구항 18**

비일시적 컴퓨터 판독 가능한 저장 매체(non-transitory computer readable storage medium)에 저장된 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은 하나 이상의 명령어들을 포함하고, 상기 명령어들은 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 실행될 때, 상기 컴퓨팅 장치로 하여금,

이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 단계;

상기 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 단계;

상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 단계;

상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 단계; 및

생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 예측하는 단계;를 수행하도록 하는, 컴퓨터 프로그램.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 발명의 실시예는 인공 신경망 기반의 객체 검출 기술로서, 청각 정보 기반의 객체 검출 기술에 관한 것이다.

### *배 경 기 술*

[0003] 영상 내에서 소리를 내는 객체 위치를 추정(Sound Source Localization)하는 기술은 멀티 모달 딥러닝(Multi-Modal Deep Learning) 분야에서 활발히 연구가 진행되고 있다. 이 기술은 로봇의 인지 기능, 음원 분리 기능, 노이즈 제어 등 다양한 어플리케이션에서 활용되고 있다.

[0004] 대부분의 객체 위치 추정 기술은 시각(Visual) 및 청각(Audio)의 두가지 정보를 활용하여 연구를 수행한다. 이는 인간의 눈에서 획득한 시각 정보와 귀로 듣는 청각 정보를 모두 활용하여 실제 세계를 바라보고 이해한다는 점을 활용한 것이다.

### *선행기술문헌*

#### *특허문헌*

[0006] (특허문헌 0001) 한국등록특허공보 제10-1812159호(2017.12.19)

(특허문헌 0002) 한국공개특허공보 제10-2021-0069462호(2021.06.11)

### *발명의 내용*

#### *해결하려는 과제*

[0007] 본 발명의 실시예는 소리를 내는 객체의 위치 추정을 위해 시각 정보에만 중점을 두었을 때 나타나는 문제점을 탈피하고, 청각 정보의 공간 정보를 활용하여 음원 위치 추적의 정확성을 높일 수 있는 공간 정보 통합 및 재귀적 집중 방식을 이용한 음원 위치 추정 방법 및 장치를 제공함에 그 목적이 있다.

## 과제의 해결 수단

[0009] 상기의 과제를 달성하기 위한 본 발명은 다음과 같은 구성을 포함한다.

[0010] 즉 본 발명의 일 실시예에 따른 음원 위치 추정 방법은 하나의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 단계, 상기 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 단계, 상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 단계, 상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 단계 및 생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 예측하는 단계를 포함한다.

[0011] 한편, 일 실시예에 따른 컴퓨팅 장치는 하나 이상의 프로세서들, 및 메모리 및 하나 이상의 프로그램들을 포함하고, 상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며, 이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 영상 획득 모듈, 상기 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 시각 위치 지역화 맵 생성 모듈, 상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 청각 위치 지역화 맵 생성 모듈, 상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 재귀적 위치 지역화 맵 생성 모듈 및 생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 생성하는 최종 음원 위치 지역화 맵 생성 모듈을 포함한다.

## 발명의 효과

[0013] 본 발명의 실시예에 의하면 소리를 내는 객체의 위치를 추정함에 있어 시각 정보에만 중점을 두었을 때 나타나는 문제점을 탈피하고, 청각 정보의 공간 정보를 활용했다는 장점을 부각시킬 수 있어 객체 위치 추정 능력을 향상시킬 수 있는 효과가 도출된다.

## 도면의 간단한 설명

[0015] 도 1 은 본 발명의 일 시시예에 따른 음원 위치 추정 방법과 기존 방법을 비교하기 위한 참고도이다.

도 2 는 본 발명의 일 실시예에 따른 음원 위치 추정 방법의 음원 위치 추정 과정을 설명하기 위한 예시도이다.

도 3 은 본 발명의 일 실시예에 따른 음원 위치 추정 장치의 구성을 나타낸 도면이다.

도 4는 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다.

## 발명을 실시하기 위한 구체적인 내용

[0016] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0017] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능

성을 배제하도록 해석되어서는 안 된다.

[0018] 또한, 제1, 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로 사용될 수 있다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0019] 개시되는 실시예에서는, 머신 러닝 또는 딥 러닝과 같은 인공 신경망을 이용하여 객체 검출을 수행하게 된다. 학습 시에는 시각 정보와 청각 정보를 모두 활용하여 객체를 검출하도록 인공 신경망을 학습할 수 있다. 그리고, 추론(테스트) 시에는 시각 정보만을 기 학습된 인공 신경망에 입력하여 객체를 검출할 수 있다.

[0021] 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명한다.

[0022] 도 1 은 본 발명의 일 시시예에 따른 음원 위치 추정 방법과 기존 방법을 비교하기 위한 참고도이다.

[0023] 도 1 에서 와 같이 기존(Existing Methods)의 소리내는 객체 즉 음원 위치 추정(Sound Source Localization) 기술은 이미지에서 관찰되는 시각 정보를 청각 정보가 따라갈 수 있도록 학습한다. 이를 통해, 청각 정보만으로 소리가 나는 객체가 무엇인지 파악할 수 있다. 또한, 시각 정보가 부족하거나 명확하지 않을 때 이를 보완하는 매개체로 사용된다.

[0024] 반면, 일 실시예에 따른 음원 위치 추정 방법(Proposed Method)은 소리를 내는 객체의 위치 추정을 위해 인간의 관점을 활용하여 이미지와 소리의 공간적인 정보를 통합하여 관심영역(ROI)를 추정하고 정확한 위치 예측을 위해 재귀적으로 집중하는 방법을 도입한다. 즉 전체적인 구조는 시각과 청각 정보의 공간적 위치가 융합된 지역화 맵을 생성한 후, 해당 지역화 맵 영역에 집중하여 다시 소리를 내는 객체의 위치에 대한 지역화 맵을 생성한다.

[0025] 이를 위해 학습 과정에서는 시청각(Audio-Visual) 정보를 모두 활용하여, 청각의 공간 정보와 시각의 공간 정보가 서로를 닮아가도록 학습한다. 결과적으로 이전 단계에서 예측한 시청각의 지역화 맵과 재귀적으로 추정한 지역화 맵을 융합하여 최종 지역화 맵을 생성한다. 이에 따라 기존의 연구 방법보다 소리 내는 객체 위치 추정 능력을 향상시킬 수 있다.

[0026] 구체적으로 도 1에서 Harp Image로부터 인코더를 통해 시각의 공간 정보(visual Spatial Feature)를 뽑아내고, 오디오 정보와 similarity Comparison을 거쳐 기존의 방식과 동일하게 이미지 맵($M_v$)을 생성할 수 있다.

[0027] 본 발명의 특징있는 양상에 따라 Harp Sound로부터 인코더를 통해 청각의 공간 정보(visual Spatial Feature)를 뽑아내고, 오디오 정보와 similarity Comparison을 거쳐 오디오 맵($M_a$)을 생성할 수 있다. 그리고 이미지 맵($M_v$)과 오디오 맵($M_a$)을 반영하여 시각과 청각 정보의 공간적 위치가 융합된 지역화 맵($M_{av}$)을 생성한다. 그리고 생성된 지역화 맵($M_{av}$)을 이용하여 청각의 공간 정보가 반영된 음원 위치를 추정할 수 있다.

[0028] 도 2 는 본 발명의 일 실시예에 따른 음원 위치 추정 방법의 음원 위치 추정 과정을 설명하기 위한 예시도이다.

[0029] 도 2 에서와 같이 일 실시예에 다른 음원 위치 추정 방법은 음원 위치를 추정함에 있어 영상의 공간 정보뿐 아니라 소리의 공간적 정보를 추가로 활용한다. 즉 소리를 내는 객체의 위치를 보다 효과적으로 찾을 수 있다. 또한 영상의 공간적 정보와 결합하여 관심 영역(ROI)을 생성하고, 이를 재귀적으로 다시 추정한다.

[0030] 즉, 시각과 청각 정보의 공간적 위치가 융합된 지역화 맵을 생성한 후, 해당 지역화 맵 영역에 집중하여 다시 소리를 내는 객체의 위치에 대한 지역화 맵을 생성한다. 그리고 이전 단계에서 예측한 시청각의 지역화 맵과 재귀적으로 추정한 지역화 맵을 융합하여 최종 지역화 맵을 생성하고, 이를 통해 소리 내는 객체의 위치를 보다 정확히 추정할 수 있다.

[0031] 도 3 은 본 발명의 일 실시예에 따른 음원 위치 추정 장치의 구성을 나타낸 도면이다.

[0032] 일 실시예에 있어서 음원 위치 추정 장치(10)는 이미지와 오디오 쌍으로 이루어진 영상 데이터를 획득하는 영상 획득 모듈(102), 영상 획득 모듈(102)에서 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성하는 시각 위치 지역화 맵 생성 모듈(112)과 상기 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성하는 청각

위치 지역화 맵 생성 모듈(114)을 포함하는 시청각 공간 통합 네트워크(110), 상기 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성하는 재귀적 위치 지역화 맵 생성 모듈(120) 및 생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 생성하는 최종 음원 위치 지역화 맵 생성 모듈(140)을 포함한다.

[0033] 영상 획득 모듈(102)은 이미지와 오디오 쌍을 받는다. 영상 획득 모듈(102)은 획득된 영상 데이터에서 이미지 촬영 시점을 기준으로 소정 시간 전후의 오디오 데이터를 획득한다. 그리고 획득된 오디오 데이터를 변환한 로그-스케일 스텍트로그램 정보를 생성한다.

[0034] 일 실시예에 있어서, 영상 획득 모듈(102)은 오디오 데이터로 이미지가 촬영된 시점으로부터 전, 후 각각 1.5초 이내의 소리 즉 총 3초간의 오디오 신호를 획득하고 이를 변환한 로그-스케일 스펙트로그램((Log-scale Spectrogram))정보를 입력받는다.

[0035] 시각 위치 지역화 맵 생성 모듈(112)은 영상 획득 모듈(102)에서 획득된 영상 데이터에서 추출된 이미지 데이터의 시각 특징점을 집중 모듈에 통과시켜 시각 위치 지역화 맵을 생성한다.

[0036] 일 실시예에 있어서 시각 위치 지역화 맵 생성 모듈(112)은 영상 획득 모듈(102)에서 획득된 영상 데이터에서 추출된 이미지 데이터를 시각 딥러닝 네트워크(Visual Encoder)에 통과시켜 시각 특징점(Image feature, $F_{v_i}$)을 도출한다.

[0037] 그리고 도출된 시각 특징점(Image feature, $F_{v_i}$)을 집중 모듈(Attention Module)에 통과시켜 시각 위치 지역화 맵(Visual Attentive Localization Map, $M_{vi}$)을 생성한다.

[0038] 이때 $l_{ai}$는 청각 특징점(Audio feature, $F_{a_i}$)을 GAP(Global Average Pooling) 해서 공간 정보는 들어있지 않은 단순 오디오 보조 정보이다.

[0039] 집중 모듈(Attention Module)은 시각 특징점(Image feature, $F_{v_i}$)과 오디오 보조 정보($l_{ai}$) 간 유사도를 측정해서 유사도 결과를 맵 형태로 표시해준다.

[0040] 예를 들어 도 2의 Visual Attentive Localization Map($M_{vi}$)과 같이 내부적으로 이미지와 오디오 간 유사도를 산출하고, 산출된 유사도가 높을 수록 붉은 색으로 표시하여 시각화할 수 있다.

[0041] 그리고 청각 위치 지역화 맵 생성 모듈(114)은 영상 획득 모듈(102)에서 획득된 오디오 데이터에서 추출된 오디오 데이터의 청각 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵을 생성한다.

[0042] 일 실시예에 있어서, 청각 위치 지역화 맵 생성 모듈(114)은 영상 획득 모듈(102)에서 획득된 영상 데이터에서 추출된 오디오 데이터를 청각 딥러닝 네트워크(Audio Encoder)에 통과시켜 청각 특징점(Audio feature, $F_{ai}$)을 도출한다.

[0043] 그리고 청각 위치 지역화 맵 생성 모듈(114)은 도출된 청각 특징점을 도출된 시각 특징점에 곱하여 오디오 중점 특징점(Audio Attentive Feature, $F_{avi}$)을 생성하고, 생성된 오디오 중점 특징점을 집중 모듈에 통과시켜 청각 위치 지역화 맵(Audio Attentive Localization Map, $M_{ai}$)을 생성한다.

[0044] 즉, 오디오 중점 특징점(Audio Attentive Feature, $F_{avi}$)과 오디오 보조 정보($l_{ai}$)를 집중 모듈(Attention Module)에 통과시켜 청각 위치 지역화 맵(Audio Attentive Localization Map, $M_{ai}$)을 생성한다.

[0045] 집중 모듈(Attention Module)은 오디오 중점 특징점(Audio Attentive Feature, $F_{avi}$)과 오디오 보조 정보($l_{ai}$) 간 유사도를 측정해서 유사도 결과를 맵 형태로 표시해준다.

[0046] 예를 들어 도 2의 Audio Attentive Localization Map($M_{ai}$)과 같이 내부적으로 오디오 중점 특징점(Audio feature, $F_{a_i}$)과 오디오($l_{ai}$)간 유사도를 산출하고, 산출된 유사도가 높을수록 붉은색으로 표시하여 시각화할 수 있다.

[0047] 그리고 재귀적 위치 지역화 맵 생성 모듈(120)은 시각 위치 지역화 맵과 청각 위치 지역화 맵을 융합하여 재귀적 위치 지역화 맵을 생성한다.

[0048] 재귀적 위치 지역화 맵 생성 모듈(120)은 시각 위치 지역화 맵 생성 모듈(112)의 시각 위치 지역화 맵(Visual Attentive Localization Map, $M_{vi}$)과 청각 위치 지역화 맵 생성 모듈(114)의 청각 위치 지역화 맵(Audio Attentive Localization Map, $M_{ai}$)을 합산하고 평균을 내서(Sum&Avg) 융합된 시청각 위치 지역화 맵(Audio-Visual Attentive Localization Map, $M_{avi}$)을 입력받는다.

[0049] 재귀적 위치 지역화 맵 생성 모듈(120)은 입력받은 시청각 위치 지역화 맵(Audio-Visual Attentive Localization Map, $M_{avi}$)과 영상 획득 모듈(102)로 입력받은 입력 이미지(Input Image, $I_{vi}$)를 산술 연산(element-wise)수행한다. 그리고 산술 연산 결과로 입력 이미지 중 음원 출력이 포함된 객체가 있을 것으로 파악되는 영역을 관심 영역으로 생성한다. 생성된 관심 영역을 식별 가능하도록 나타낸 집중 이미지(Attentive Input Image, $I_{v_t}^{att}$ )를 도출해낼 수 있다. 이때 식별 가능하도록 나타내는 것은 해당 영역을 밝게 표시하는 것일 수 있다.

[0050] 재귀적 위치 지역화 맵 생성 모듈(120)은 도출된 집중 이미지(Attentive Input Image, $I_{v_t}^{att}$ )를 시각 딥러닝 네트워크(Visual Encoder)에 통과시켜 집중 시각 특징점(Image feature, $F_{v_i}^{att}$ )을 도출한다.

[0051] 그리고 도출된 재귀적 시각 특징점(Image feature, $F_{v_i}^{att}$ )을 집중 모듈(Attention Module)에 통과시켜 재귀적 위치 지역화 맵(Visual Attentive Localization Map, $M_{v_i}^{att}$ )을 생성한다.

[0052] 이때 $l_{ai}$는 청각 특징점(Audio feature, $F_{a_i}$)을 GAP(Global Average Pooling) 해서 공간 정보는 들어있지 않은 단순 오디오 보조 정보이다.

[0053] 집중 모듈(Attention Module)은 재귀적 시각 특징점(Image feature, $F_{v_i}^{att}$ )과 오디오 보조 정보($l_{ai}$) 간 유사도를 측정해서 유사도 결과를 재귀적 위치 지역화 맵(Visual Attentive Localization Map, $M_{v_i}^{att}$) 형태로 표시해준다.

[0054] 예를 들어 도 2의 재귀적 위치 지역화 맵(Visual Attentive Localization Map, $M_{v_i}^{att}$)과 같이 내부적으로 이미지와 오디오 간 유사도를 산출하고, 산출된 유사도가 높을 수록 붉게 가시적으로 표시해준다.

[0055] 최종 음원 위치 지역화 맵 생성 모듈(140)은 시각 위치 지역화 맵 생성 모듈(112)에서 생성된 시각 위치 지역화 맵, 청각 위치 지역화 맵 생성 모듈(114)에서 생성된 청각 위치 지역화 맵, 재귀적 위치 지역화 맵 생성 모듈(120)에서 생성된 재귀적 위치 지역화맵을 융합하여 최종 음원 위치 지역화 맵을 생성한다.

[0056] 결과적으로 3개의 위치 지역화 맵(시각 위치 지역화 맵, 청각 위치 지역화 맵, 재귀적 위치 지역화 맵)을 각각의 가중치를 고려하여 융합하여 최종 위치 지역화 맵(Final Localization Map, $M_{final_i}$)을 예측할 수 있다. 예를 들어, 최종 위치 지역화 맵(Final Localization Map, $M_{final_i}$)에서 각 픽셀별로 0~1 사이의 값들 중 소정의 기준 이상으로 1에 인접한 부분을 바운딩 처리하여 음원 객체로 인식할 수 있다. 그러나 이에 한정되는 것은 아니다.

[0057] 손실 함수 도출 모듈(130)은 집중 모듈 통과시 생성된 손실함수, 시청각 맞춤 손실 함수, 공간 영역 조정 손실 함수를 합산하여 최종 학습 손실함수를 도출한다.

[0058] 이를 식으로 나타내면 수학식 1과 같다.

*수학식 1*

[0059] $$L_{total}=L_{SSL}+\lambda_1 L_{avpm}+\lambda_2 L_{sra}$$

[0060] 여기서 $L_{SSL}$은 집중 모듈을 통과할 때 생성된 손실함수이고, 이는 시청각 짝을 맞춰주는 대조 손실 함수

(Contrastive Loss)이다. $L_{SSL}$은 집중 모듈에서 생성된 손실 함수들의 합일 수 있다.

[0061] 상기 수학식 1을 바탕으로 소리의 공간 정보의 활용이 가능하여 강인한 음원 위치 추정이 가능해진다.

[0062] $L_{avpm}$은 시청각 맞춤 손실함수(Audio-Visual Pair Matching Loss)로 이는 변형된 삼중항 손실(Triplet Loss) 함수로, 기준점(anchor)이 되는 데이터와 쌍을 이루는(positive) 데이터는 가까운 거리에 위치하고, 쌍을 이루지 않는(negative) 데이터는 먼 거리에 위치하도록 만들어준다.

[0063] $L_{avpm}$는 수학식 2와 같다.

*수학식 2*

[0064] $$L_{avpm}=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1(j\neq i)}^{N}T(l_{v_i}^{att},l_{a_i},l_{a_j})+T(l_{v_i}^{att},l_{v_i},l_{v_j})$$

[0065] 먼저, 시청각 맞춤 손실함수(Audio-Visual Pair Matching Loss)를 도출하기 위해 청각 특징점과 시각 특징점이 집중된 시각 특징점을 따라갈 수 있는 손실 함수를 수학식 3과 같이 제안한다.

*수학식 3*

[0066] $$D(\alpha,\beta)=||(\alpha-\beta)/\tau||_2^2$$

[0067] $$T(l_{v_i}^{att},l_{a_i},l_{a_j})=D(l_{vi}^{att},l_{a_i})+\max(\delta-D(l_{vi}^{att},l_{a_j}),0),$$

[0068] $$T(l_{v_i}^{att},l_{v_i},l_{v_j})=D(l_{vi}^{att},l_{v_i})+\max(\delta-D(l_{vi}^{att},l_{v_j}),0)$$

[0069] 여기서 δ와 τ는 값 조정을 위한 초매개변수(hyper parameter)이다. D(α,β)는 α와 β사이의 유클리드 거리(Euclidean distance)를 의미한다.

[0070] 그리고 $l_{v_i}^{att}$ 는 재귀적 시각 특징점의 GAP(Global Average Pooling) 값, $l_{v_i}$는 시각 특징점의 GAP(Global Average Pooling) 값, $l_{a_i}$는 청각 특징점의 GAP(Global Average Pooling) 값, D(α,β)는 α와 β사이의 유클리드 거리(Euclidean distance)이다.

[0071] T(*a, p, n*)는 기준점(anchor)이 되는 데이터 a와 쌍을 이루는(positive) 데이터 *p*의 거리는 작아지도록, 쌍을 이루지 않는(negative) 데이터 *n*은 거리가 멀어지도록 구성한다.

[0072] 이를 토대로 $T(l_{v_i}^{att},l_{a_i},l_{a_j})$는 $l_{vi}^{att}$와 쌍을 이루는 오디오 데이터 $l_{a_i}$와는 가깝게 구성되고, 쌍을 이루지 않는 오디오 데이터 $l_{a_j}$와는 멀어지도록 구성된다. 여기서 $l_v,l_a,l_v^{att}$ 는 각각 시각 특징점, 청각 특징점, 재귀적 시각 특징점을 GAP(Global Average Pooling)을 한 것이다. 즉 기준이 되는 i 제품의 이미지 데이터와 오디오 데이터의 거리는 가깝게 하고, 다른 i 제품의 오디오 데이터와는 거리가 멀어지도록 학습한다.

[0073] 비슷하게 $T(l_{v_i}^{att},l_{v_i},l_{v_j})$도 이미지 데이터로 구성한다. 시청각 맞춤 손실 함수 $L_{avpm}$(Audio-Visual Pair Matching Loss)는 이를 바탕으로 구성될 수 있다.

[0074] GAP(Global Average Pooling)는 각 픽셀들의 값들을 평균으로 내서 하나로 만드는 것으로 feature 수를 줄일 수 있다. 이는 비주얼 정보와 오디오 정보가 동일한 객체를 인지하도록 하기 위한 것이다.

[0075] 그리고 공간 영역 조정 손실 함수 $L_{sra}$(Spatial Region Alignment Loss)는 청각 특징점이 공간적인 정보를 가질 수 있도록 집중된 시각 특징점을 닮아가도록 학습한다.

[0076] 이를 위한 공간 영역 조정 손실 함수 $L_{sra}$(Spatial Region Alignment Loss)는 수학식 4와 같이 도출될 수 있다.

*수학식 4*

[0077] $$L_{sra}=\frac{1}{N}\sum_{i=1}^{N}D_{KL}(\hat{G}_{v_i}^{att}\parallel\hat{G}_{a_i})$$

[0078] 여기서 $D_{KL}$은 쿨백-라이블러 발산(Kullback-Leibler divergence)으로, 이를 줄이는 것을 통해 청각 특징점이 시각 특징점의 공간 영역을 닮아가도록 학습할 수 있다.

[0079] 여기서 N은 학습시 입력되는 영상 데이터 수, $\hat{G}_{v}^{att}$ 는 재귀적 시각 특징점을 채널 요소 평균(Channel-wise Average Pooling)을 한 후 1차원 벡터화(Flatten)한것, $\hat{G}_{a}$는 청각 특징점을 채널 요소 평균(Channel-wise Average Pooling)을 한 후 1차원 벡터화(Flatten)한 것이다.

[0080] 도 2에서 오디오 특징점인 $F_{ai}$의 채널 요소 평균(Channel-wise Average Pooling)을 통해 with-height 공간적 위치 정보만 남겨줌으로써 $\hat{F}_{a_i}$를 만들고, 이를 다시 1차원 벡터화(Flatten)하여 $\hat{G}_{ai}$를 생성한다.

[0081] 그리고 집중 시각 특징점(Image feature, $F_{v_i}^{att}$ )의 채널 요소 평균(Channel-wise Average Pooling)을 통해 with-height 공간적 위치 정보만 남겨줌으로써 $\hat{F}_{v_i}^{att}$ 를 만들고, 이를 다시1차원 벡터화(Flatten)하여 $\hat{G}_{vi}^{att}$ 를 생성한다.

[0082] 그리고 $\hat{G}_{ai}$가 $\hat{G}_{vi}^{att}$ 를 따라가도록 학습을 수행한다.

[0083] 일 실시예에 따르면 기존의 시각 정보를 중점으로 소리를 내는 객체 위치를 검출하는 기술과 달리, 청각 정보도 공간 정보로 함께 활용하여 위치를 검출할 수 있다. 이에 따라 소리를 내지 않는 객체를 함께 검출하여 잠재적으로 사고를 유발할 수 있는 상황에 대한 대처가 불가능 하다는 기존의 시각 정보가 가지고 있는 한계를 벗어날 수 있는 장점이 있다.

[0084] 학습 과정에서 영상의 시청각 정보를 활용함에 의해 멀티모달 정보를 활용할 수 있고, 단일 시각 및 청각 정보의 정보만 전달받는 기존의 기술보다 더욱 발전된 기술을 제시할 수 있다. 또한, 객체를 인식할 때 여러 번 반복하여 인식함으로써 네트워크를 한번만 거쳐서 위치를 추정하는 기존의 기술보다 더욱 정확한 위치 추정이 가능하다는 효과가 있다.

[0085] 도 4는 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0086] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 음원 위치 추정 장치(100)일 수 있다.

[0087] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0088] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메

모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0089] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0090] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.

[0092] 이상에서 본 발명의 대표적인 실시예들을 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

## *부호의 설명*

[0094] 10 : 컴퓨팅 장치 14 : 프로세서

16 : 컴퓨터 판독 가능 저장 매체 18 : 통신 버스

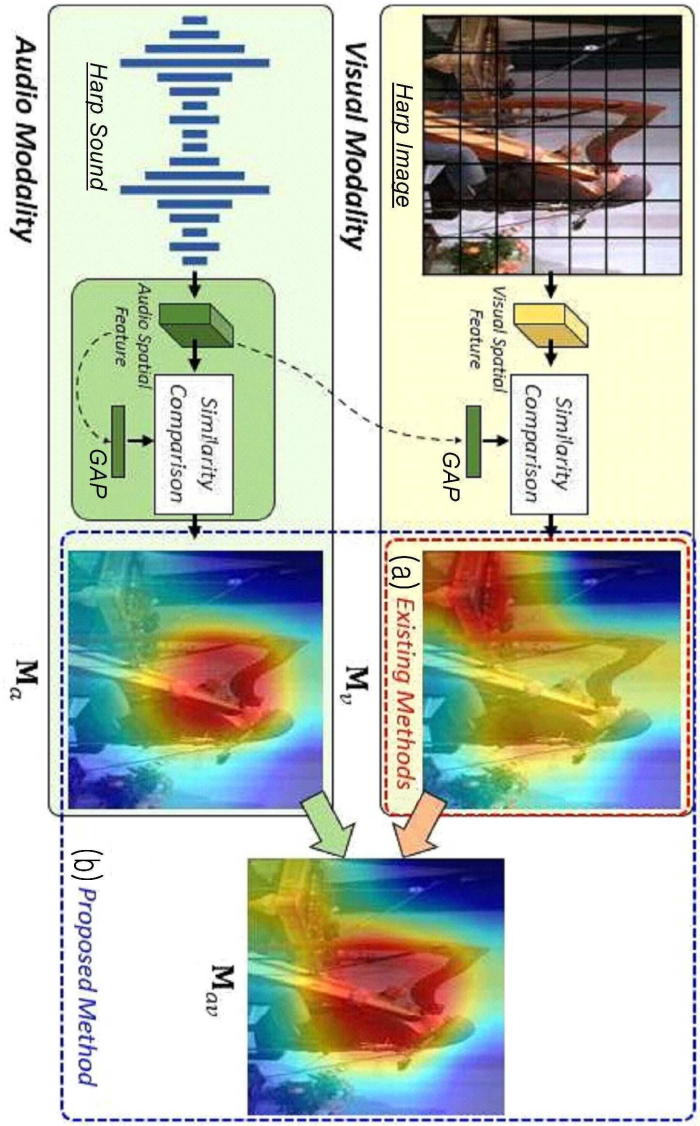
20 : 프로그램 22 : 입출력 인터페이스

24 : 입출력 장치 26 : 네트워크 통신 인터페이스

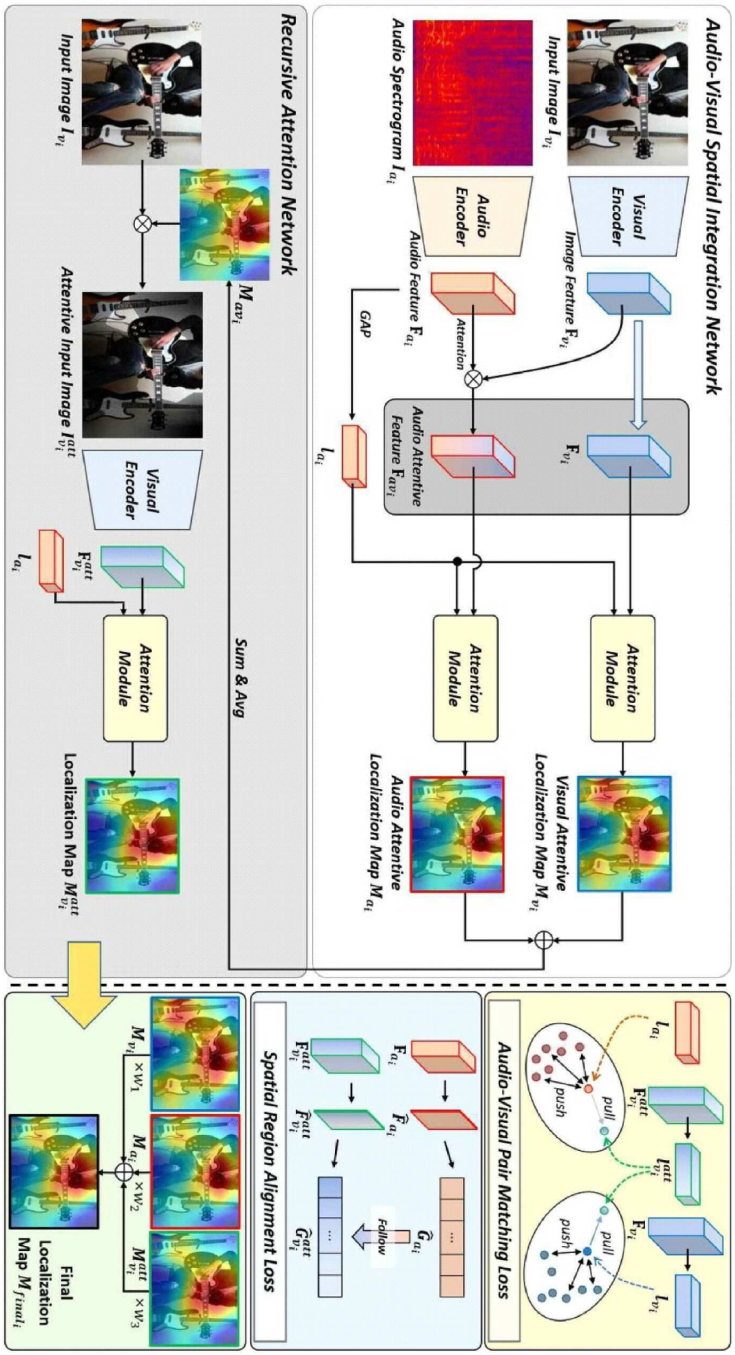
100 : 음원 위치 추정 장치

도면

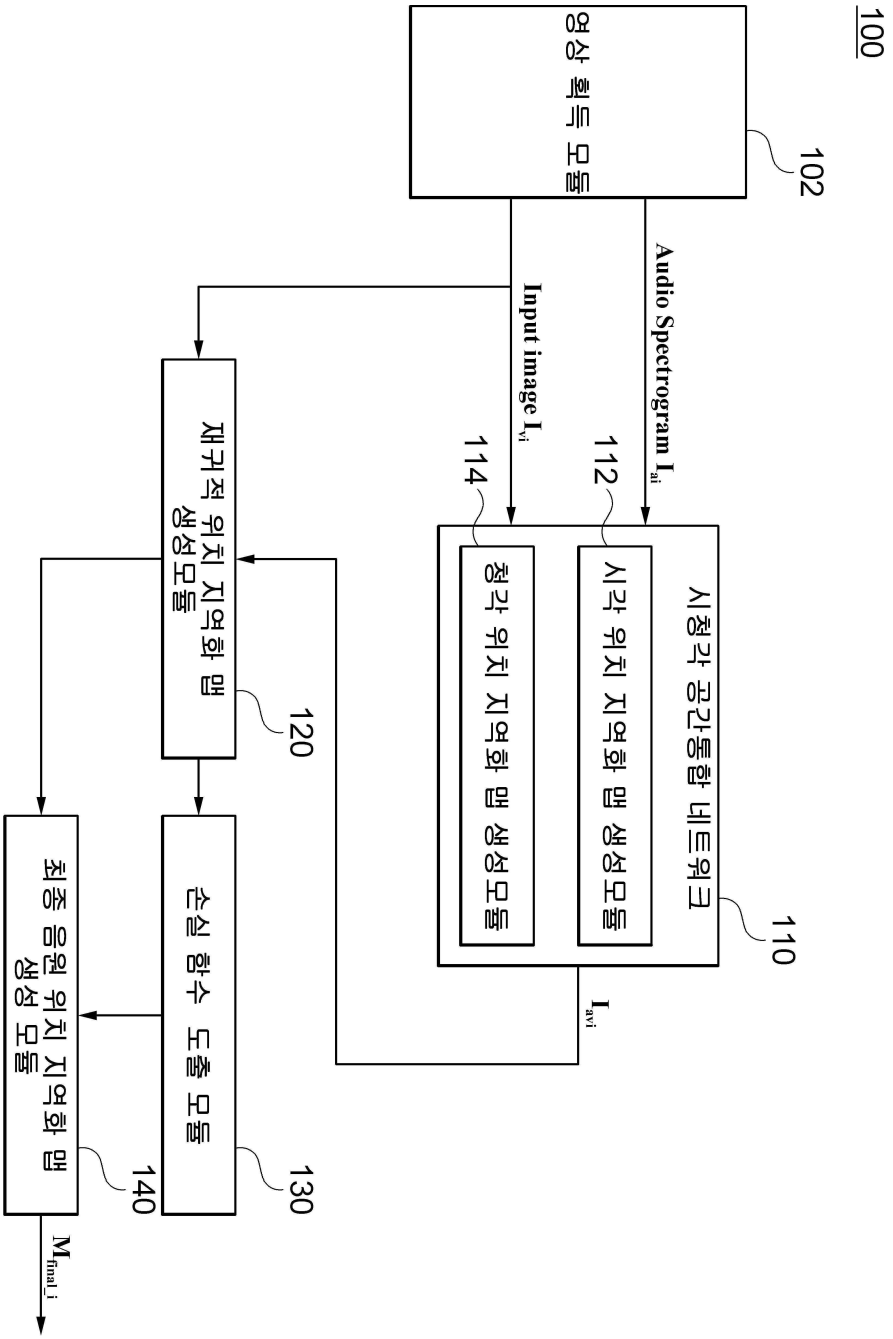
도면1


Audio Modality
Harp Sound
Audio Spatial Feature
Similarity Comparison
GAP
$\mathbf{M}_a$
Visual Modality
Harp Image
Visual Spatial Feature
Similarity Comparison
GAP
$\mathbf{M}_v$
(a) Existing Methods
$\mathbf{M}_{av}$
(b) Proposed Method

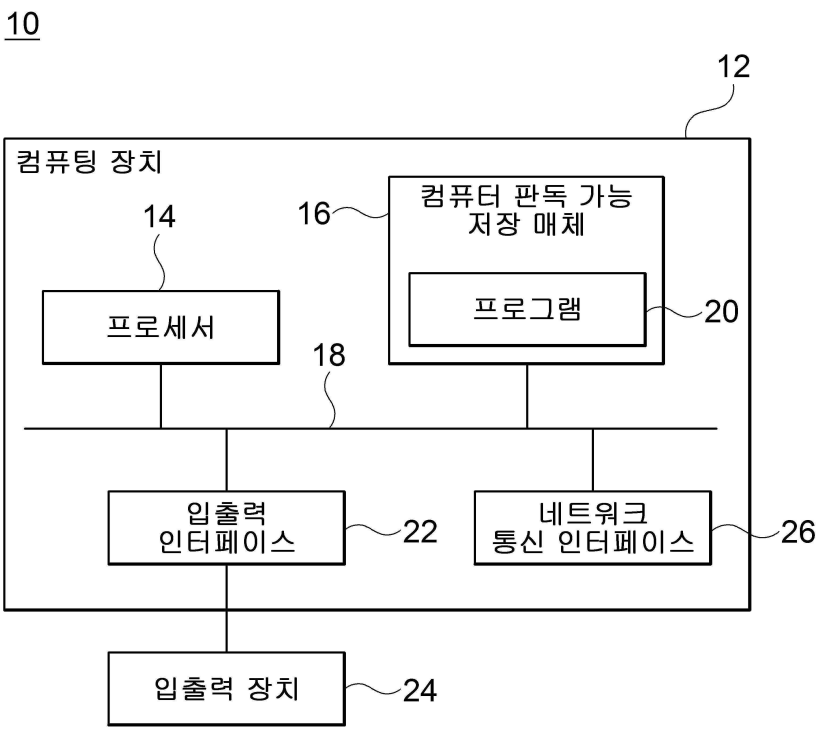
도면2


Audio-Visual Spatial Integration Network
Input Image $I_{v_i}$
Visual Encoder
Image Feature $F_{v_i}$
Audio Spectrogram $I_{a_i}$
Audio Encoder
Audio Feature $F_{a_i}$
Attention
GAP
$l_{a_i}$
$F_{v_i}$
Audio Attentive Feature $F_{av_i}$
Attention Module
Attention Module
Visual Attentive Localization Map $M_{v_i}$
Audio Attentive Localization Map $M_{a_i}$
Recursive Attention Network
Input Image $I_{v_i}$
$M_{av_i}$
Attentive Input Image $I_{v_i}^{att}$
Visual Encoder
$F_{v_i}^{att}$
$l_{a_i}$
Attention Module
Sum & Avg
Localization Map $M_{v_i}^{att}$
Audio-Visual Pair Matching Loss
$l_{a_i}$
$F_{v_i}^{att}$
$l_{v_i}^{att}$
$F_{v_i}$
$l_{v_i}$
pull
push
Spatial Region Alignment Loss
$F_{a_i}$
$\hat{F}_{a_i}$
$F_{v_i}^{att}$
$\hat{F}_{v_i}^{att}$
$\hat{G}_{a_i}$
Follow
$\hat{G}_{v_i}^{att}$
$M_{v_i}$
$\times w_1$
$M_{a_i}$
$\times w_2$
$M_{t_i}^{att}$
$\times w_3$
Final Localization Map $M_{final_i}$

도면3


100
영상 획득 모듈
102
Audio Spectrogram $I_{ai}$
Input image $I_{vi}$
112
114
시청각 공간통합 네트워크
110
시각 위치 지역화 맵 생성모듈
청각 위치 지역화 맵 생성모듈
$I_{avi}$
재귀적 위치 지역화 맵 생성모듈
120
손실 함수 도출 모듈
130
최종 음원 위치 지역화 맵 생성 모듈
140
$M_{final_i}$

도면4


10
12
컴퓨팅 장치
14
프로세서
16
컴퓨터 판독 가능
저장 매체
프로그램
20
18
입출력
인터페이스
22
네트워크
통신 인터페이스
26
입출력 장치
24