(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년08월26일
(11) 등록번호 10-2148607
(24) 등록일자 2020년08월20일

(51) 국제특허분류(Int. Cl.)
H04N 21/43 (2011.01)
(52) CPC특허분류
H04N 21/4307 (2020.08)
H04N 21/4394 (2013.01)
(21) 출원번호 10-2019-0090937
(22) 출원일자 2019년07월26일
심사청구일자 2019년07월26일
(56) 선행기술조사문헌
논문:Yapeng Tian et al.,*
KR1020180038937 A
KR101900237 B1
KR101731461 B1
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
손광훈
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C129호(신촌동)
이지영
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C129호(신촌동)
(74) 대리인
민영준

전체 청구항 수 : 총 8 항 심사관 : 정윤석

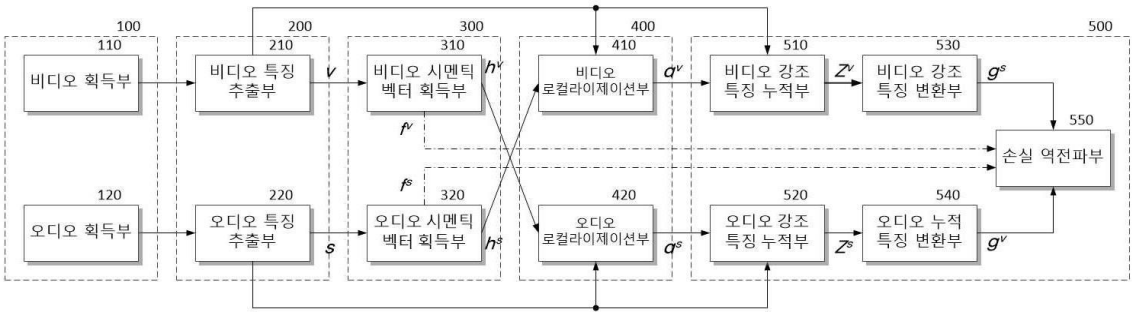
(54) 발명의 명칭 **오디오-비디오 정합 영역 탐지 장치 및 방법**

(57) 요 약

본 발명은 미리 학습된 패턴 추정 방식에 따라 비디오 데이터와 오디오 데이터 각각에서 특징을 추출하여 비디오 특징맵과 오디오 특징맵을 획득하는 특징맵 획득부, 비디오 특징맵과 오디오 특징맵을 기지정된 동일한 차원을 갖는 비디오 변환 특징맵과 오디오 변환 특징맵으로 변환하고, 미리 학습된 패턴 추정 방식에 따라 비디오 변환

(뒷면에 계속)

대 표 도



특징맵과 오디오 변환 특징맵 각각의 특징을 추출하여 비디오 시멘틱 벡터와 오디오 시멘틱 벡터를 획득하는 시멘틱 벡터 획득부 및 비디오 특징맵과 오디오 시멘틱 벡터를 기지정된 방식으로 결합하여 비디오 특징맵에서 오디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 비디오 강조맵을 획득하고, 오디오 특징맵과 비디오 시멘틱 벡터를 기지정된 방식으로 결합하여 오디오 특징맵에서 비디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 오디오 강조맵을 획득하는 로컬라이제이션부를 포함하여, 비디오에서 인식된 객체에 대응하는 오디오 구간을 검출하거나 오디오로부터 비디오의 대응하는 객체 영역을 검출할 수 있도록 하는 오디오-비디오 정합 영역 탐지 장치 및 방법을 제공할 수 있다.

(52) CPC특허분류

H04N 21/44008 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2019070646
부처명	과학기술정보통신부
연구관리전문기관	한국연구재단
연구사업명	원천기술개발사업
연구과제명	(2세부)딥러닝 기반 의미론적 상황 이해 원천기술 연구 (2단계)(1/2)
기 여 율	1/1
주관기관	연세대학교 산학협력단
연구기간	2019.06.01 ~ 2020.03.31

명 세 서

청구범위

청구항 1

미리 학습된 패턴 추정 방식에 따라 비디오 데이터와 오디오 데이터 각각에서 특징을 추출하여 비디오 특징맵과 오디오 특징맵을 획득하는 특징맵 획득부;

상기 비디오 특징맵과 상기 오디오 특징맵을 기지정된 동일한 차원을 갖는 비디오 변환 특징맵과 오디오 변환 특징맵으로 변환하고, 미리 학습된 패턴 추정 방식에 따라 상기 비디오 변환 특징맵과 상기 오디오 변환 특징맵 각각의 특징을 추출하여 비디오 시멘틱 벡터와 오디오 시멘틱 벡터를 획득하는 시멘틱 벡터 획득부; 및

상기 비디오 특징맵과 상기 오디오 시멘틱 벡터를 기지정된 방식으로 결합하여 상기 비디오 특징맵에서 상기 오디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 비디오 강조맵을 획득하고, 상기 오디오 특징맵과 상기 비디오 시멘틱 벡터를 기지정된 방식으로 결합하여 상기 오디오 특징맵에서 비디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 오디오 강조맵을 획득하는 로컬라이제이션부; 를 포함하되,

상기 특징맵 획득부와 상기 시멘틱 벡터 획득부를 학습시키기 위한 학습부를 더 포함하고,

상기 학습부는

상기 비디오 강조맵의 원소들과 상기 비디오 특징맵의 곱을 누적하여 비디오 누적 강조 특징맵을 획득하는 비디오 강조 특징 누적부;

상기 오디오 강조맵의 원소들과 상기 오디오 특징맵의 곱을 누적하여 오디오 누적 강조 특징맵을 획득하는 오디오 강조 특징 누적부;

학습 시에 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부와 함께 패턴 추정 방식이 학습되어 상기 비디오 누적 강조 특징맵의 특징을 추출하여 오디오 강조 특징 벡터를 획득하는 오디오 강조 특징 변환부;

학습 시에 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부와 함께 패턴 추정 방식이 학습되어 상기 오디오 누적 강조 특징맵의 특징을 추출하여 비디오 강조 특징 벡터를 획득하는 비디오 강조 특징 변환부; 및

상기 비디오 변환 특징맵과 상기 비디오 강조 특징 벡터 사이의 차와 상기 오디오 변환 특징맵과 상기 오디오 강조 특징 벡터 사이의 차를 합하여 손실을 계산하고, 계산된 손실을 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부, 상기 오디오 강조 특징 변환부 및 상기 비디오 강조 특징 변환부로 역전파하는 손실 역전파부; 를 포함하는 오디오-비디오 정합 영역 탐지 장치.

청구항 2

제1 항에 있어서, 상기 시멘틱 벡터 획득부는

상기 비디오 특징맵을 인가받아 상기 비디오 변환 특징맵으로 변환하는 비디오 특징 차원 변환부;

상기 오디오 특징맵을 인가받아 상기 오디오 변환 특징맵으로 변환하는 오디오 특징 차원 변환부;

상기 오디오 시멘틱 벡터가 반영된 상기 비디오 강조맵에 기반하여 패턴 추정 방식이 미리 학습되어 상기 비디오 변환 특징맵으로부터 상기 비디오 시멘틱 벡터를 추출하는 비디오 시멘틱 벡터 추출부; 및

상기 비디오 시멘틱 벡터가 반영된 상기 오디오 강조맵에 기반하여 패턴 추정 방식이 미리 학습되어 상기 오디오 변환 특징맵으로부터 상기 오디오 시멘틱 벡터를 추출하는 오디오 시멘틱 벡터 추출부; 를 포함하는 오디오-비디오 정합 영역 탐지 장치.

청구항 3

제1 항에 있어서, 상기 로컬라이제이션부는

상기 비디오 특징맵의 부분 행렬들 각각에 대한 전치 행렬과 상기 오디오 시멘틱 벡터를 행렬 곱하여 비디오 강

조 벡터를 획득하는 비디오 강조 벡터 획득부;

상기 오디오 특징맵의 부분 행렬들 각각에 대한 전치 행렬과 상기 비디오 시멘틱 벡터를 행렬 곱하여 오디오 강조 벡터를 획득하는 오디오 강조 벡터 획득부; 및

상기 비디오 강조 벡터와 상기 비디오 강조 벡터를 인가받아 기지정된 방식으로 정규화하여 상기 비디오 강조맵과 상기 오디오 강조맵을 획득하는 강조 벡터 정규화부; 를 포함하는 오디오-비디오 정합 영역 탐지 장치.

청구항 4

제3 항에 있어서, 상기 강조 벡터 정규화부는

소프트맥스 함수에 의한 확률에 기반하여 상기 비디오 강조 벡터와 상기 비디오 강조 벡터를 정규화하는 오디오-비디오 정합 영역 탐지 장치.

청구항 5

삭제

청구항 6

미리 학습된 패턴 추정 방식에 따라 비디오 데이터와 오디오 데이터 각각에서 특징을 추출하여 비디오 특징맵과 오디오 특징맵을 획득하는 단계;

상기 비디오 특징맵과 상기 오디오 특징맵을 기지정된 동일한 차원을 갖는 비디오 변환 특징맵과 오디오 변환 특징맵으로 변환하는 단계;

미리 학습된 패턴 추정 방식에 따라 상기 비디오 변환 특징맵과 상기 오디오 변환 특징맵 각각의 특징을 추출하여 비디오 시멘틱 벡터와 오디오 시멘틱 벡터를 획득하는 단계;

상기 비디오 특징맵과 상기 오디오 시멘틱 벡터를 기지정된 방식으로 결합하고 상기 오디오 특징맵과 상기 비디오 시멘틱 벡터를 기지정된 방식으로 결합하여, 상기 비디오 특징맵에서 상기 오디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 비디오 강조맵과 상기 오디오 특징맵에서 비디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 오디오 강조맵을 획득하는 단계; 를 포함하되,

상기 오디오 특징맵 및 상기 비디오 특징맵을 획득하는 단계와 상기 비디오 시멘틱 벡터와 상기 오디오 시멘틱 벡터를 획득하는 단계를 학습시키기 위한 학습 단계를 더 포함하고,

상기 학습 단계는

상기 비디오 강조맵의 원소들과 상기 비디오 특징맵의 곱을 누적하여 비디오 누적 강조 특징맵을 획득하는 단계;

상기 오디오 강조맵의 원소들과 상기 오디오 특징맵의 곱을 누적하여 오디오 누적 강조 특징맵을 획득하는 단계;

학습 시에 패턴 추정 방식이 학습되어 상기 비디오 누적 강조 특징맵의 특징을 추출하여 오디오 강조 특징 벡터를 획득하는 단계;

학습 시에 패턴 추정 방식이 학습되어 상기 오디오 누적 강조 특징맵의 특징을 추출하여 비디오 강조 특징 벡터를 획득하는 단계; 및

상기 비디오 변환 특징맵과 상기 비디오 강조 특징 벡터 사이의 차와 상기 오디오 변환 특징맵과 상기 오디오 강조 특징 벡터 사이의 차를 합하여 손실을 계산하고, 계산된 손실을 역전파하는 단계; 를 포함하는 오디오-비디오 정합 영역 탐지 방법.

청구항 7

제6 항에 있어서, 상기 시멘틱 벡터를 획득하는 단계는

상기 오디오 시멘틱 벡터가 반영된 상기 비디오 강조맵에 기반하여 패턴 추정 방식이 미리 학습되어 상기 비디

오 변환 특징맵으로부터 상기 비디오 시멘틱 벡터를 추출하는 단계; 및

상기 비디오 시멘틱 벡터가 반영된 상기 오디오 강조맵에 기반하여 패턴 추정 방식이 미리 학습되어 상기 오디오 변환 특징맵으로부터 상기 오디오 시멘틱 벡터를 추출하는 단계; 를 포함하는 오디오-비디오 정합 영역 탐지 방법.

청구항 8

제6 항에 있어서, 상기 강조맵을 획득하는 단계는

상기 비디오 특징맵의 부분 행렬들 각각에 대한 전치 행렬과 상기 오디오 시멘틱 벡터를 행렬 곱하여 비디오 강조 벡터를 획득하는 단계;

상기 오디오 특징맵의 부분 행렬들 각각에 대한 전치 행렬과 상기 비디오 시멘틱 벡터를 행렬 곱하여 오디오 강조 벡터를 획득하는 단계; 및

상기 비디오 강조 벡터와 상기 비디오 강조 벡터를 인가받아 상기 비디오 강조맵과 상기 오디오 강조맵을 획득하기 위해 기지정된 방식으로 정규화하는 단계; 를 포함하는 오디오-비디오 정합 영역 탐지 방법.

청구항 9

제8 항에 있어서, 상기 정규화하는 단계는

소프트맥스 함수에 의한 확률에 기반하여 상기 비디오 강조 벡터와 상기 비디오 강조 벡터를 정규화하는 오디오-비디오 정합 영역 탐지 방법.

청구항 10

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 오디오-비디오 정합 영역 탐지 장치 및 방법에 관한 것으로, 오디오로부터 비디오에서 대응하는 객체 영역을 탐지하거나, 비디오의 객체에 대응하는 오디오 구간을 탐지할 수 있는 오디오-비디오 정합 영역 탐지 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근 개인의 멀티미디어 방송의 활성화에 따라 간단한 멀티미디어 편집 기술에 대한 요구가 급증하고 있다. 이러한 멀티미디어 편집 시에는 영상에 적합한 다양한 효과음이 포함하거나 효과음에 대응하는 영상을 제공해야 하는 경우가 빈번하게 발생한다. 다양한 객체에 대한 비디오나 오디오와 같은 멀티미디어 데이터는 기존의 검색 기법을 통해 용이하게 획득될 수 있다. 그리고 획득된 영상 및 음향에서 지정된 특정 객체의 이미지 또는 음향이 포함된 영역 및 구간을 검출하기 위해 객체 영역 로컬라이제이션(object area localization) 기법이 제시된바 있다.

[0003] 즉 주어진 비디오나 오디오에 포함된 객체를 인식하거나, 비디오나 오디오에서 주어진 객체에 대응하는 구간을 탐지하기 위한 다양한 기술이 연구되어 왔다. 그러나, 객체의 다양성과 복잡한 배경 등과 같은 여러 이유로 인해 성능의 제약이 있다.

[0004] 이에 최근에는 딥 러닝(Deep learning) 기법으로 학습된 인공 신경망(artificial neural network)을 이용하여 비디오 또는 오디오에서 객체 영역을 추출하는 로컬라이제이션을 수행하기 위한 다양한 연구가 진행되었으며, 딥 러닝 기법을 이용함에 의해 비디오에 대한 객체 영역 로컬라이제이션 작업의 성능이 크게 향상되었다.

[0005] 그러나 기존의 연구에서 객체 영역 로컬라이제이션은 주로 비디오와 오디오 각각에 대해 개별적으로 주어진 색인어에 대한 객체 영역을 검출하거나, 비디오나 오디오에 포함된 객체를 인지하여 객체에 대한 태그를 추출하는 방식으로 수행되었다. 따라서 동일 객체에 대한 유의어 등에 취약하여 요구하는 객체에 대해 정확하게 지정된 태그가 제시되지 않는다면 검출 성능이 크게 저하되는 문제가 있다. 또한 비디오로부터 객체의 오디오 구간을 직접 추출하거나, 오디오로부터 비디오에 포함된 객체 영역을 직접 추출하지 못한다는 한계가 있다. 뿐만 아니

라 색인어와 같은 문자어를 이용하는 경우, 단순히 객체를 지정하는 방식이므로 객체의 존재 여부나 전체 윤곽을 판별하기에는 용이하지만, 비디오에서 음향이 발생되는 영역을 나타내거나, 오디오에서 특정 객체의 음향이 포함된 구간을 추출하거나 주변 잡음을 제거하기에는 부적합하다.

[0006] 특히 멀티미디어 편집 시에는 비디오와 오디오 사이에 동기화(synchronization)가 이루어지지 않는 경우가 빈번하게 발생한다. 그러나 기존에는 비디오와 오디오 사이의 정합을 제공하기 위한 기준을 제시할 수 없어, 멀티미디어를 편집하고자 하는 편집자가 매번 수작업으로 동기화를 수행해야 하는 번거로움이 있었다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국 등록 특허 제10-1900237호 (2018.09.13 등록)

발명의 내용

해결하려는 과제

[0008] 본 발명의 목적은 오디오로부터 비디오의 대응하는 객체 영역을 검출할 수 있는 오디오-비디오 정합 영역 탐지 장치 및 방법을 제공하는데 있다.

[0009] 본 발명의 다른 목적은 비디오의 객체를 인식하고, 인식된 객체에 대응하는 오디오 구간을 검출할 수 있는 오디오-비디오 정합 영역 탐지 장치 및 방법을 제공하는데 있다.

[0010] 본 발명의 또 다른 목적은 오디오와 비디오의 동기화를 위한 구간을 자동으로 검출할 수 있도록 하는 오디오-비디오 정합 영역 탐지 장치 및 방법을 제공하는데 있다.

[0011] 본 발명의 또 다른 목적은 주석 처리된 학습용 데이터를 요구하지 않고, 비디오와 오디오 사이의 상호 시멘틱 특징을 기반으로 자기 지도 학습 방식으로 학습되어 오디오와 비디오에서 동일 객체에 대한 구간 검출을 위한 학습을 동시에 수행할 수 있는 오디오-비디오 정합 영역 탐지 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0012] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 오디오-비디오 정합 영역 탐지 장치는 미리 학습된 패턴 추정 방식에 따라 비디오 데이터와 오디오 데이터 각각에서 특징을 추출하여 비디오 특징맵과 오디오 특징맵을 획득하는 특징맵 획득부; 상기 비디오 특징맵과 상기 오디오 특징맵을 기지정된 동일한 차원을 갖는 비디오 변환 특징맵과 오디오 변환 특징맵으로 변환하고, 미리 학습된 패턴 추정 방식에 따라 상기 비디오 변환 특징맵과 상기 오디오 변환 특징맵 각각의 특징을 추출하여 비디오 시멘틱 벡터와 오디오 시멘틱 벡터를 획득하는 시멘틱 벡터 획득부; 및 상기 비디오 특징맵과 상기 오디오 시멘틱 벡터를 기지정된 방식으로 결합하여 상기 비디오 특징맵에서 상기 오디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 비디오 강조맵을 획득하고, 상기 오디오 특징맵과 상기 비디오 시멘틱 벡터를 기지정된 방식으로 결합하여 상기 오디오 특징맵에서 비디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 오디오 강조맵을 획득하는 로컬라이제이션부; 를 포함한다.

[0013] 상기 시멘틱 벡터 획득부는 상기 비디오 특징맵을 인가받아 상기 비디오 변환 특징맵으로 변환하는 비디오 특징 차원 변환부; 상기 오디오 특징맵을 인가받아 상기 오디오 변환 특징맵으로 변환하는 오디오 특징 차원 변환부; 상기 오디오 시멘틱 벡터가 반영된 상기 비디오 강조맵에 기반하여 패턴 추정 방식이 미리 학습되어 상기 비디오 변환 특징맵으로부터 상기 비디오 시멘틱 벡터를 추출하는 비디오 시멘틱 벡터 추출부; 및 상기 비디오 시멘틱 벡터가 반영된 상기 오디오 강조맵에 기반하여 패턴 추정 방식이 미리 학습되어 상기 오디오 변환 특징맵으로부터 상기 오디오 시멘틱 벡터를 추출하는 오디오 시멘틱 벡터 추출부; 를 포함할 수 있다.

[0014] 상기 로컬라이제이션부는 상기 비디오 특징맵의 부분 행렬들 각각에 대한 전치 행렬과 상기 오디오 시멘틱 벡터를 행렬 곱하여 비디오 강조 벡터를 획득하는 비디오 강조 벡터 획득부; 상기 오디오 특징맵의 부분 행렬들 각각에 대한 전치 행렬과 상기 비디오 시멘틱 벡터를 행렬 곱하여 오디오 강조 벡터를 획득하는 오디오 강조 벡터 획득부; 및 상기 비디오 강조 벡터와 상기 비디오 강조 벡터를 인가받아 기지정된 방식으로 정규화하여 상기 비디오 강조맵과 상기 오디오 강조맵을 획득하는 강조 벡터 정규화부; 를 포함할 수 있다.

[0015] 상기 강조 벡터 정규화부는 소프트맥스 함수에 의한 확률에 기반하여 상기 비디오 강조 벡터와 상기 비디오 강조 벡터를 정규화할 수 있다.

[0016] 상기 오디오-비디오 정합 영역 탐지 장치는 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부를 학습시키기 위한 학습부를 더 포함할 수 있다.

[0017] 상기 학습부는 상기 비디오 강조맵의 원소들과 상기 비디오 특징맵의 곱을 누적하여 비디오 누적 강조 특징맵을 획득하는 비디오 강조 특징 누적부; 상기 오디오 강조맵의 원소들과 상기 오디오 특징맵의 곱을 누적하여 오디오 누적 강조 특징맵을 획득하는 오디오 강조 특징 누적부; 학습 시에 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부와 함께 패턴 추정 방식이 학습되어 상기 비디오 누적 강조 특징맵의 특징을 추출하여 오디오 강조 특징 벡터를 획득하는 오디오 강조 특징 변환부; 학습 시에 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부와 함께 패턴 추정 방식이 학습되어 상기 오디오 누적 강조 특징맵의 특징을 추출하여 비디오 강조 특징 벡터를 획득하는 비디오 강조 특징 변환부; 및 상기 비디오 변환 특징맵과 상기 비디오 강조 특징 벡터 사이의 차와 상기 오디오 변환 특징맵과 상기 오디오 강조 특징 벡터 사이의 차를 합하여 손실을 계산하고, 계산된 손실을 상기 특징맵 획득부와 상기 시멘틱 벡터 획득부, 상기 오디오 강조 특징 변환부 및 상기 비디오 강조 특징 변환부로 역전파하는 손실 역전파부; 를 포함할 수 있다.

[0018] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 오디오-비디오 정합 영역 탐지 방법은 미리 학습된 패턴 추정 방식에 따라 비디오 데이터와 오디오 데이터 각각에서 특징을 추출하여 비디오 특징맵과 오디오 특징맵을 획득하는 단계; 상기 비디오 특징맵과 상기 오디오 특징맵을 기지정된 동일한 차원을 갖는 비디오 변환 특징맵과 오디오 변환 특징맵으로 변환하는 단계; 미리 학습된 패턴 추정 방식에 따라 상기 비디오 변환 특징맵과 상기 오디오 변환 특징맵 각각의 특징을 추출하여 비디오 시멘틱 벡터와 오디오 시멘틱 벡터를 획득하는 단계; 상기 비디오 특징맵과 상기 오디오 시멘틱 벡터를 기지정된 방식으로 결합하고 상기 오디오 특징맵과 상기 비디오 시멘틱 벡터를 기지정된 방식으로 결합하여, 상기 비디오 특징맵에서 상기 오디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 비디오 강조맵과 상기 오디오 특징맵에서 비디오 시멘틱 벡터에 따른 위치별 강조 세기를 나타내는 오디오 강조맵을 획득하는 단계; 를 포함한다.

발명의 효과

[0019] 따라서, 본 발명의 실시예에 따른 오디오-비디오 정합 영역 탐지 장치 및 방법은 비디오의 객체를 인식하고, 인식된 객체에 대응하는 오디오 구간을 검출하거나 오디오로부터 비디오의 대응하는 객체 영역을 검출할 수 있도록 한다. 그러므로 객체에 대한 정확한 태그를 알지 못하는 상태에서도 멀티미디어에서 비디오 또는 오디오 중 하나로부터 나머지 하나의 대응하는 영역 또는 구간을 추출하여 오디오와 비디오의 동기화를 위한 구간을 자동으로 검출할 수 있다. 뿐만 아니라 오디오와 비디오에 대한 학습을 동시에 수행할 수 있어 학습의 효율성을 크게 향상 시킬 수 있으며, 오디오와 비디오 사이의 상호 시멘틱 상관 관계에 따른 자가 지도 학습 방식으로 학습을 수행하여 별도로 주석된 학습용 데이터를 필요로 하지 않는다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일 실시예에 따른 오디오-비디오 정합 영역 탐지 장치의 개략적 구조를 나타낸다.

도 2는 도 1의 시멘틱 벡터 획득부의 상세 구성을 나타낸다.

도 3은 도 1의 로컬라이제이션부의 상세 구성을 나타낸다.

도 4는 도 1의 오디오-비디오 정합 영역 탐지 장치의 각 구성별 동작을 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시예에 따른 오디오-비디오 정합 영역 탐지 방법을 나타낸다.

도 6 및 도 7은 본 발명의 오디오-비디오 정합 영역 탐지 방법을 이용하여 멀티미디어에서 객체의 음향에 대응하는 영역을 추출한 결과를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0022] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러

나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0023] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0024] 도 1은 본 발명의 일 실시예에 따른 오디오-비디오 정합 영역 탐지 장치의 개략적 구조를 나타내고, 도 2는 도 1의 시멘틱 벡터 획득부의 상세 구성을 나타내며, 도 3은 도 1의 로컬라이제이션부의 상세 구성을 나타낸다. 그리고 도 4은 도 1의 오디오-비디오 정합 영역 탐지 장치의 각 구성별 동작을 설명하기 위한 도면이다.

[0025] 도 1 및 도 4를 참조하면, 본 실시예에 따른 오디오-비디오 정합 영역 탐지 장치는 멀티미디어 획득부(100), 특징맵 획득부(200), 시멘틱 벡터 획득부(300) 및 로컬라이제이션부(400)를 포함한다.

[0026] 멀티미디어 획득부(100)는 동일 객체에 대해 대응하는 영역을 탐지하고자 하는 멀티미디어 데이터를 획득한다. 여기서 멀티미디어 데이터는 비디오 데이터와 오디오 데이터를 포함할 수 있다. 그리고 멀티미디어 데이터는 동영상과 같이 비디오 데이터와 오디오 데이터가 동일 장소에서 함께 획득된 데이터일 수 있으나, 비디오와 오디오가 개별적으로 획득된 데이터이어도 무방하다. 즉 서로 다른 시간과 장소에서 유사한 객체에 대해 획득된 비디오 데이터와 오디오 데이터일 수 있다.

[0027] 멀티미디어 획득부(100)는 직접 멀티미디어 데이터를 생성하는 비디오 카메라 와 같은 멀티미디어 장치로 구현될 수도 있으나, 유/무선 네트워크를 통해 멀티미디어 데이터를 전송받는 통신부 또는 이전 획득한 멀티미디어 데이터를 저장하는 저장 장치 등으로 구현될 수도 있다.

[0028] 멀티미디어 획득부(100)는 비디오 데이터를 획득하는 비디오 획득부(110)와 오디오 데이터를 획득하는 오디오 획득부(120)를 포함한다. 만일 동영상과 같이 비디오 데이터와 오디오 데이터가 하나의 파일로 통합된 멀티미디어 데이터가 획득된 경우, 비디오 획득부(110)와 오디오 획득부(120)는 멀티미디어 데이터에서 비디오 데이터와 오디오 데이터를 각각 분리하여 획득할 수도 있다.

[0029] 또한 비디오 획득부(110)는 비디오 데이터가 연속되는 다수의 프레임을 포함하는 경우, 다수 프레임 각각을 구분하여 개별 프레임의 이미지를 추출하여 비디오 데이터로 획득할 수도 있다.

[0030] 특징맵 획득부(200)는 멀티미디어 획득부(100)에서 획득된 비디오 데이터와 오디오 데이터 각각의 특징을 추출하여, 비디오 특징맵(V)과 오디오 특징맵(S)을 획득한다. 특징맵 획득부(200)는 비디오 데이터의 특징을 추출하는 비디오 특징 추출부(210)와 오디오 데이터의 특징을 추출하는 오디오 특징 추출부(220)를 포함한다.

[0031] 비디오 특징 추출부(210)와 오디오 특징 추출부(220) 각각은 패턴 추정 방식이 각각 미리 학습된 인공 신경망을 포함하여 구현되어 학습된 패턴 추정 방식에 따라 비디오 데이터의 특징과 오디오 데이터의 특징을 추출하여, 비디오 특징맵(V)과 오디오 특징맵(S)을 획득한다. 여기서 비디오 특징 추출부(210)는 2차원의 비디오 데이터로부터 실수값(R)을 원소로 높이(H), 폭(W) 및 깊이(D)를 갖는 3차원 행렬(또는 벡터) 형태($V \in R^{H\times W\times D}$)의 비디오 특징맵(V)을 획득하고, 오디오 특징 추출부(220)는 1차원의 오디오 데이터로부터 실수값(R)을 원소로 강도(M)와 깊이(D)를 갖는 2차원 행렬(또는 벡터) 형태($S \in R^{M\times D}$)의 오디오 특징맵(S)을 획득할 수 있다. 그러나 비디오 특징맵(V)과 오디오 특징맵(S)의 차원 및 크기는 다양하게 조절될 수 있다.

[0032] 비디오 데이터와 오디오 데이터로부터 비디오 특징맵(V)과 오디오 특징맵(S)을 획득하는 인공 신경망은 다양하게 공개되어 있다. 여기서는 일예로 도 3에 도시된 바와 같이, 컨볼루션 신경망은 적어도 하나의 컨볼루션 레이어(conv)와 적어도 하나의 풀링 레이어(pool)를 포함하는 컨볼루션 신경망(Convolutional Neural Networks)으로 구현되는 것으로 가정한다.

[0033] 또한 본 실시예에서 비디오 특징 추출부(210)와 오디오 특징 추출부(220)는 후술하는 학습부(500)에서 계산된 손실이 역전파되어 특징맵(V, S)을 획득하기 위한 학습이 수행될 수 있다. 그러나 비디오 특징맵(V)과 오디오 특징맵(S)을 획득하기 위해 패턴 추정 방식이 미리 학습된 다양한 인공 신경망이 공개되어 있으므로, 경우에 따라서는 이러한 학습된 인공 신경망을 이용할 수도 있다.

[0034] 한편, 시멘틱 벡터 획득부(300)는 미리 학습된 패턴 추정 방식에 따라 특징맵 획득부(200)에서 획득된 비디오

특징맵(V)과 오디오 특징맵(S) 각각으로부터 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)를 획득한다. 여기서 비디오 시멘틱 벡터(h^v)는 비디오 특징맵(V)의 패턴으로부터 추출되는 특징 벡터로서 비디오 데이터의 공간 축에서 객체의 위치를 나타내기 위한 벡터이고, 오디오 시멘틱 벡터(h^s)는 오디오 특징맵(S)의 패턴으로부터 추출되는 특징 벡터로서 오디오 데이터의 시간 축에서 객체의 위치를 나타내기 위한 벡터이다.

[0035] 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)는 본 실시예의 오디오-비디오 정합 영역 탐지 장치가 비디오 데이터와 오디오 데이터서 공통으로 포함된 유사 객체의 영역을 판별할 수 있도록 하기 위해 획득되는 정보로서, 비디오 특징과 오디오 특징 사이에 상호 공통 객체 영역을 검출하기 위해 획득된다. 즉 본 실시예에서 비디오 시멘틱 벡터(h^v)는 오디오 데이터에서 객체의 음향이 발생된 영역을 검출하기 위해 획득되는 벡터이고, 오디오 시멘틱 벡터(h^s)는 비디오 데이터에서 객체가 나타나는 영역을 검출하기 위해 획득되는 벡터이다.

[0036] 일반적으로 멀티미디어 획득부(100)에서 획득되는 비디오 데이터에는 특정 객체뿐만 아니라 주변 배경과 함께 다양한 객체가 포함될 수 있다. 또한 오디오 데이터에도 특정 객체에서 발생된 음향 이외에도 주변의 다양한 음향이 함께 포함될 수 있다. 따라서 비디오 데이터와 오디오 데이터에 공통의 객체에 대한 정보가 포함되어 있더라도, 비디오 데이터와 오디오 데이터 각각에서 개별적으로 획득된 비디오 특징맵(V)과 오디오 특징맵(S)에서 공통의 객체가 포함된 영역을 검출하기 용이하지 않다. 다만 특징맵 획득부(200)의 비디오 특징 추출부(210)와 오디오 특징 추출부(220)가 정상적으로 학습된 상태라면, 특징맵 비디오 특징맵(V)과 오디오 특징맵(S)에는 공통 객체에 대한 특징이 포함될 것으로 추정할 수 있다. 이에 시멘틱 벡터 획득부(300)는 학습된 패턴 추정 방식에 따라 비디오 특징맵(V)과 오디오 특징맵(S)의 객체의 위치를 나타내는 주요 특징에 대한 패턴을 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)로 추출한다.

[0037] 본 실시예에서는 시멘틱 벡터 획득부(300)가 서로 다른 종류의 데이터인 비디오 데이터와 오디오 데이터 각각으로부터 획득된 특징맵에서 의미적론(semantic)으로 동일한 객체에 대한 영역을 탐지하기 위해 획득하는 특징 벡터를 추출하므로, 추출된 특징 벡터를 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)로 정의하였다.

[0038] 시멘틱 벡터 획득부(300)는 비디오 특징맵(V)으로부터 비디오 시멘틱 벡터(h^s)를 획득하는 비디오 시멘틱 벡터 획득부(310)와 오디오 특징맵(S)으로부터 오디오 시멘틱 벡터(h^s)를 획득하는 오디오 시멘틱 벡터 획득부(320)를 포함한다.

[0039] 도 2를 참조하면, 비디오 시멘틱 벡터 획득부(310)는 비디오 특징 차원 변환부(311)와 비디오 시멘틱 벡터 추출부(312)를 포함하고, 오디오 시멘틱 벡터 획득부(320)는 오디오 특징 차원 변환부(321)와 오디오 시멘틱 벡터 추출부(322)를 포함할 수 있다.

[0040] 비디오 특징 차원 변환부(311)와 오디오 특징 차원 변환부(321)는 특징맵 획득부(200)에서 인가되는 비디오 특징맵(V)과 오디오 특징맵(S)의 크기를 동일하게 일치시키기 위해 포함되는 구성이다. 상기한 바와 같이, 특징맵 획득부(200)에서 획득되는 비디오 특징맵(V)은 3차원(H × W × D) 행렬 형태로 획득되는 반면, 오디오 특징맵(S)은 2차원(M × D) 행렬 형태로 획득되므로, 서로 차원 및 크기가 상이하며, 이로 인해 공통의 객체에 대한 특징을 추출하기 어려우며, 추출하더라도 상호 적용이 용이하지 않다.

[0041] 이에 비디오 특징 차원 변환부(311)와 오디오 특징 차원 변환부(321)는 비디오 특징맵(V)과 오디오 특징맵(S)을 기지정된 방식으로 동일한 1차원 행렬로 변환한다.

[0042] 일예로 비디오 특징 차원 변환부(311)는 3차원(H × W × D) 행렬로 구성되는 비디오 특징맵(V)에서 깊이(D)를 기준으로 높이(H) 및 폭(W)에 대한 원소들의 평균값을 획득함으로써, 1차원으로 차원 변환하여 비디오 변환 특징맵($f^v \in R^D$)을 획득할 수 있다.

[0043] 오디오 특징 차원 변환부(321) 또한 유사하게 2차원(M × D) 행렬인 오디오 특징맵(S)을 인가받아 1차원 행렬로 변환하여 오디오 변환 특징맵($f^s \in R^D$)을 획득할 수 있다.

[0044] 비디오 시멘틱 벡터 추출부(312)와 오디오 시멘틱 벡터 추출부(322)는 비디오 변환 특징맵(f^v)과 오디오 변환

특징맵(f^s)을 인가받고, 각각 패턴 추정 방식이 미리 학습된 인공 신경망으로 구현되어 인가된 비디오 변환 특징맵(f^v)과 오디오 변환 특징맵(f^s)의 특징 벡터인 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)를 획득한다.

[0045] 로컬라이제이션부(400)는 특징맵 획득부(200)에서 획득된 비디오 특징맵(V) 및 오디오 특징맵(S)과 시멘틱 벡터 획득부(300)에서 획득된 비디오 시멘틱 벡터(h^v) 및 오디오 시멘틱 벡터(h^s)를 이용하여 비디오 강조맵(a^v)과 오디오 강조맵(a^s)을 획득한다.

[0046] 도 1 및 도 3을 참조하면, 로컬라이제이션부(400)는 비디오 로컬라이제이션부(410)와 오디오 로컬라이제이션부(420)를 포함할 수 있다. 그리고 비디오 로컬라이제이션부(410)는 비디오 강조 벡터 획득부(411)와 비디오 강조 벡터 정규화부(412)를 포함하고, 오디오 로컬라이제이션부(420)는 오디오 강조 벡터 획득부(451)와 오디오 강조 벡터 정규화부(422)를 포함할 수 잇다.

[0047] 비디오 강조 벡터 획득부(411)는 비디오 특징맵(V)과 오디오 시멘틱 벡터(h^s)를 인가받아 수학식 1에 따라 비디오 강조 벡터(c^v)를 획득한다.

수학식 1

[0048] $$c^v = v^T h^s$$

[0049] 여기서 v^T는 3차원인 비디오 특징맵(V)의 부분 행렬(v)에 대한 전치 행렬을 의미한다.

[0050] 비디오 강조 벡터(c^v)는 비디오 특징맵(V)에서 오디오 시멘틱 벡터(h^s)에 따른 위치별 강조 세기를 나타낸다.

[0051] 그리고 비디오 강조 벡터 정규화부(412)는 비디오 강조 벡터(c^v)를 인가받고, 기지정된 방식으로 비디오 강조 벡터(c^v)를 정규화하여 비디오 강조맵(a^v)을 획득한다. 이때 비디오 강조 벡터 정규화부(412)는 일예로 소프트맥스 함수(softmax function)에 의한 확률에 따라 수학식 2와 같이 비디오 강조 벡터(c^v)를 정규화할 수 있다.

수학식 2

[0052] $$a^v = \frac{\exp(c^v)}{\sum \exp(c^v)}$$

[0053] 한편, 오디오 강조 벡터 획득부(421)는 오디오 특징맵(S)과 비디오 시멘틱 벡터(h^v)를 인가받아 수학식 3에 따라 오디오 특징맵(S)에서 비디오 시멘틱 벡터(h^v)에 따른 위치별 강조 세기를 나타내는 오디오 강조 벡터(c^s)를 획득한다.

수학식 3

[0054] $$c^s = s^T h^v$$

[0055] 여기서 s^T는 2차원인 오디오 특징맵(S)의 부분 행렬(s)에 대한 전치 행렬을 의미한다.

[0056] 그리고 오디오 강조 벡터 정규화부(422)는 비디오 강조 벡터 정규화부(412)와 유사하게 오디오 강조 벡터(c^s)를 소프트맥스 함수에 의한 확률에 따라 수학식 4와 같이 오디오 강조 벡터(c^s)를 정규화하여 오디오 강조맵(a^s)을

획득할 수 있다.

수학식 4

[0057]
$$a^s = \frac{\exp(c^s)}{\sum \exp(c^s)}$$

[0058] 로컬라이제이션부(400)에서 획득되는 비디오 강조맵(a^v)과 오디오 강조맵(a^s)은 비디오 데이터의 위치별 오디오 데이터의 특징에 대응하는 강도 및 오디오 데이터의 위치별 비디오 데이터의 특징에 대응하는 강도를 나타내는 행렬이다.

[0059] 따라서 비디오 강조맵(a^v)과 오디오 강조맵(a^s)은 비디오 데이터와 오디오 데이터 각각에서 서로 대응하는 객체 영역을 표현할 수 있다.

[0060] 도 3에서는 이해를 위해 비디오 강조 벡터 정규화부(412)와 오디오 강조 벡터 정규화부(422)를 구분하였으나, 비디오 강조 벡터 정규화부(412)와 오디오 강조 벡터 정규화부(422)는 강조 벡터 정규화부로 통합될 수 있다.

[0061] 한편, 오디오-비디오 정합 영역 탐지 장치가 비디오 데이터와 오디오 데이터에서 공통된 객체에 대한 영역을 바르게 검출하기 위해서는 특징맵 획득부(200) 및 시멘틱 벡터 획득부(300)의 인공 신경망이 미리 학습되어야 하며, 이에 본 실시예에 따른 오디오-비디오 정합 영역 탐지 장치는 특징맵 획득부(200) 및 시멘틱 벡터 획득부(300)의 인공 신경망이 학습시키기 위한 학습부(500)를 더 포함할 수 있다.

[0062] 학습부(500)는 오디오-비디오 정합 영역 탐지 장치의 학습 과정에 필요한 구성으로 오디오-비디오 정합 영역 탐지 장치가 학습된 이후, 실제 운용 시에는 제외될 수 있다.

[0063] 다시 도 1을 참조하면, 학습부(500)는 비디오 강조 특징 누적부(510), 오디오 강조 특징 누적부(520), 오디오 강조 특징 변환부(530), 비디오 강조 특징 변환부(540) 및 손실 역전파부(550)를 포함할 수 있다.

[0064] 우선 비디오 강조 특징 누적부(510)는 비디오 로컬라이제이션부(410)에서 획득된 비디오 강조맵(a^v)과 비디오 특징 추출부(210)에서 획득된 비디오 특징맵(V)을 인가받고, 인가된 비디오 강조맵(a^v)의 각 원소(a_i^v)와 비디오 특징맵(V)의 곱을 누적하여 비디오 누적 강조 특징맵(Z^v)을 수학식 5와 같이 획득한다. 여기서 i(i ∈ {1, ..., HW})는 공간적 위치를 나타낸다.

[0065] 그리고 오디오 강조 특징 누적부(520)는 오디오 로컬라이제이션부(420)에서 획득된 오디오 강조맵(a^s)과 오디오 특징 추출부(210)에서 획득된 오디오 특징맵(S)을 인가받고, 인가된 오디오 강조맵(a^s)의 각 원소(a_j^s)와 오디오 특징맵(S)의 곱을 누적하여 오디오 누적 강조 특징맵(Z^s)을 수학식 6과 같이 획득한다. 여기서 j(j ∈ {1, ..., M})는 시간적 위치를 나타낸다.

[0066] 한편, 비디오 누적 특징 변환부(530)와 오디오 누적 특징 변환부(540)는 각각 인공 신경망으로 구현된다. 인공 신경망으로 구현되는 비디오 누적 특징 변환부(530)와 오디오 누적 특징 변환부(540)는 특징맵 획득부(200) 및 시멘틱 벡터 획득부(300)의 학습 시에 함께 학습이 수행되어, 오디오-비디오 정합 영역 탐지 장치가 별도의 주석이 포함되지 않은 멀티미디어 데이터를 인가받아 자가 학습이 수행되도록 한다.

[0067] 비디오 강조 특징 변환부(530)는 비디오 강조 특징 누적부(510)에서 인가되는 비디오 누적 강조 특징맵(Z^v)에서 특징을 추출하여, 오디오 특징 추출부(220)와 오디오 시멘틱 벡터 획득부(320)를 학습시키기 위한 오디오 강조 특징 벡터(g^s)를 획득한다.

[0068] 그리고 오디오 강조 특징 변환부(540)는 오디오 강조 특징 누적부(520)에서 인가되는 오디오 누적 강조 특징맵(Z^s)에서 특징을 추출하여, 비디오 특징 추출부(210)와 비디오 시멘틱 벡터 획득부(310)를 학습시키기 위한 비

디오 강조 특징 벡터(g^v)를 획득한다.

[0069] 만일 학습이 정상적으로 수행된 상태라면, 비디오 데이터에서 객체의 특징을 추출한 비디오 특징맵(V)의 차원을 변환한 비디오 변환 특징맵(f^v)과 비디오 시멘틱 벡터(h^v)에 의해 강조된 오디오 누적 강조 특징맵(Z^s)의 특징을 추출한 비디오 강조 특징 벡터(g^v)는 유사하게 나타나야 한다($f^v \fallingdotseq g^v$). 또한 오디오 데이터에서 객체의 특징을 추출한 오디오 특징맵(S)의 차원을 변환한 오디오 변환 특징맵(f^s)과 오디오 시멘틱 벡터(h^s)에 의해 강조된 비디오 누적 강조 특징맵(Z^v)의 특징을 추출한 오디오 강조 특징 벡터(g^s)는 유사하게 나타나야 한다($f^s \fallingdotseq g^s$).

[0070] 이에 비디오 강조 특징 변환부(530)는 오디오 특징 추출부(220)와 오디오 시멘틱 벡터 획득부(320)의 오디오 시멘틱 벡터 추출부(322)를 자기 지도 학습 시킬 수 있도록 손실을 계산하기 위한 오디오 강조 특징 벡터(g^s)를 획득하고, 오디오 강조 특징 변환부(540)는 비디오 특징 추출부(210)와 비디오 시멘틱 벡터 획득부(310)의 비디오 시멘틱 벡터 추출부(312)를 자기 지도 학습 시킬 수 있도록 손실을 계산하기 위한 비디오 강조 특징 벡터(g^v)를 획득한다.

[0071] 한편 손실 역전파부(550)는 시멘틱 벡터 획득부(300)에서 획득된 비디오 변환 특징맵(f^v)과 비디오 강조 특징 벡터(g^v) 사이의 차와 오디오 변환 특징맵(f^s)과 오디오 강조 특징 벡터(g^s) 사이의 차에 기반하여, 수학식 5와 같이 손실(L)을 계산한다.

수학식 5

[0072] $$\mathcal{L} = ||f^v - g^v||_2 + \lambda||f^s - g^s||_2$$

[0073] 여기서 λ는 비디오 손실과 오디오 손실 사이의 중요도를 조절하기 위한 매개 변수이고, $\| \ \|_2$ 는 L_2-norm 함수이다.

[0074] 그리고 손실 역전파부(550)는 계산된 손실(L)을 특징맵 획득부(200) 및 시멘틱 벡터 획득부(300)와 함께 비디오 강조 특징 변환부(530) 및 오디오 강조 특징 변환부(540)로 역전파하여 학습시킨다.

[0075] 즉 본 실시예의 오디오-비디오 정합 영역 탐지 장치에서는 특징맵 획득부(200) 및 시멘틱 벡터 획득부(300)와 함께 비디오 강조 특징 변환부(530) 및 오디오 강조 특징 변환부(540)가 자가 지도 학습이 수행된다. 이때, 손실 역전파부(550)는 계산된 손실(L)이 기지정된 문턱 손실(L_{th}) 이하이거나, 반복 횟수가 기지정된 반복 학습 횟수에 도달하면 학습을 종료할 수 있다.

[0076] 다만 상기한 바와 같이, 특징맵 획득부(200)는 미리 학습이 수행된 인공 신경망을 적용할 수 있으며, 이 경우, 특징맵 획득부(200)로는 손실을 역전파하여 학습시키지 않을 수도 있다. 또한 특징맵 획득부(200)가 이미 학습된 상태일지라도 손실(L)을 역전파하여 추가적인 학습이 수행되도록 할 수도 있다. 이는 비록 추가 학습을 수행하는 경우일지라도, 특징맵 획득부(200)가 이전에 객체 탐지를 위한 학습이 수행된 상태라면 학습 속도를 향상시킬 수 있기 때문이다.

[0077] 한편 상기한 바와 같이 본 실시예에 따른 오디오-비디오 정합 영역 탐지 장치는 서로 다른 장소나 시간 등에서 별도로 획득된 비디오 데이터나 오디오 데이터에 대해서도 동일 객체에 대한 영역을 용이하게 탐지할 수 있다. 다만 학습을 수행하는 경우에는 가급적 서로 대응하는 비디오 데이터와 오디오 데이터가 이용되는 것이 바람직하며, 이에 동일 장소에서 동일 시간에 비디오 데이터와 오디오 데이터가 함께 획득된 동영상이 이용되는 것이 바람직하다.

[0078] 도 5는 본 발명의 일 실시예에 따른 오디오-비디오 정합 영역 탐지 방법을 나타낸다.

[0079] 도 1 내지 도 4를 참조하여, 도 5의 오디오-비디오 정합 영역 탐지 방법을 설명하면, 오디오-비디오 정합 영역 탐지 방법은 오디오-비디오 정합 영역 탐지 단계(S10) 및 학습 단계(S20)를 포함할 수 있다.

[0080] 여기서 학습 단계(S20)는 오디오-비디오 정합 영역 탐지 방법의 운용 이전 학습 시에만 이용되므로, 실제 운용 시에는 생략될 수 있다.

[0081] 오디오-비디오 정합 영역 탐지 단계(S10)를 살펴보면, 우선 정합 영역이 탐지되어야 하는 비디오 데이터와 오디오 데이터를 획득한다(S11). 그리고 미리 학습된 패턴 추정 방식에 따라 획득된 비디오 데이터와 오디오 데이터 각각에서 특징을 추출하여 비디오 특징맵(V)과 오디오 특징맵(S)을 획득한다(S12).

[0082] 여기서 획득된 비디오 특징맵(V)과 오디오 특징맵(S)의 차원이 서로 상이하므로, 비디오 특징맵(V)과 오디오 특징맵(S)이 동일 차원이 되도록 1차원으로 변환하여, 비디오 변환 특징맵(f^v)과 오디오 변환 특징맵(f^s)을 획득한다(S13).

[0083] 이후, 공통의 특징을 추출하기 위해 미리 학습된 패턴 추정 방식에 따라 비디오 변환 특징맵(f^v)과 오디오 변환 특징맵(f^s) 각각으로부터 공통의 특징을 추출하여 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)를 획득한다(S14).

[0084] 비디오 시멘틱 벡터(h^v)와 오디오 시멘틱 벡터(h^s)가 획득되면, 비디오 특징맵(V)과 오디오 시멘틱 벡터(h^s)를 기지정된 방식으로 결합하고 정규화하여 오디오 데이터에 대응하는 비디오 데이터의 공간적 강조 영역을 나타내는 비디오 강조맵(a^v)을 획득하고, 오디오 특징맵(S)과 비디오 시멘틱 벡터(h^v)를 기지정된 방식으로 결합하고 정규화하여 비디오 데이터에 대응하는 오디오 데이터의 시간적 강조 영역을 나타내는 오디오 강조맵(a^s)을 획득한다(S15).

[0085] 한편 학습 단계(S20)에서는 학습을 위한 멀티미디어 데이터, 즉 학습을 위해 획득된 비디오 데이터와 오디오 데이터로부터 획득된 비디오 강조맵(a^v)과 오디오 강조맵(a^s)의 각 원소와 비디오 특징맵(V) 및 오디오 특징맵(S)의 곱을 누적하여, 비디오 누적 강조 특징맵(Z^v)과 오디오 누적 강조 특징맵(Z^s)을 획득한다(S21).

[0086] 그리고 획득된 비디오 누적 강조 특징맵(Z^v)의 특징을 추정하여, 오디오 강조 특징 벡터(g^s)으로 변환하고, 오디오 누적 강조 특징맵(Z^s)의 특징을 추정하여, 비디오 강조 특징 벡터(g^v)으로 변환한다(S22).

[0087] 이후 획득된 비디오 변환 특징맵(f^v)과 비디오 강조 특징 벡터(g^v) 사이의 차와 오디오 변환 특징맵(f^s)과 오디오 강조 특징 벡터(g^s) 사이의 차에 기반하여, 수학식 5와 같이 손실(L)을 계산한다(S23).

[0088] 손실(L)이 계산되면, 계산된 손실(L)을 역전파하여 학습을 수행한다(S24).

[0089] 여기서 오디오-비디오 정합 영역 탐지 방법의 학습은 오디오-비디오 정합 영역 탐지 단계(S10) 및 학습 단계(S20) 전체를 반복하여 수행되며, 반복 수행 횟수가 기지정된 반복 학습 횟수에 도달하거나, 는 계산된 손실(L)이 기지정된 문턱 손실(L_{th}) 이하이면 종료될 수 있다.

[0090] 도 6 및 도 7은 본 발명의 오디오-비디오 정합 영역 탐지 방법을 이용하여 멀티미디어에서 객체의 음향에 대응하는 영역을 추출한 결과를 나타낸다.

[0091] 본 실시예에 따른 오디오-비디오 정합 영역 탐지 장치 및 방법은 도 6 및 도 7 각각의 (a) 내지 (c)에 도시된 바와 같이 공통의 특정 객체가 포함된 비디오 데이터와 오디오 데이터가 주어지면, (d) 내지 (f)와 같이 오디오 데이터에 대응하는 비디오 데이터의 영역을 검출하거나 비디오 데이터에 대응하는 오디오 데이터의 구간을 정확하게 검출할 수 있다. 비록 도 6 및 도 7에서는 시각적 표현의 용이성에 따라 오디오 데이터에 대응하는 비디오 데이터의 강조 영역만을 표시하였으나, 비디오 데이터에 따른 오디오 데이터의 구간도 정확하게 추출될 수 있다.

[0092] 본 발명에 따른 방법은 컴퓨터에서 실행 시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0093] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0094] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0095] 100: 멀티미디어 획득부 200: 특징맵 획득부

300: 시멘틱 벡터 획득부 400: 로컬라이제이션부

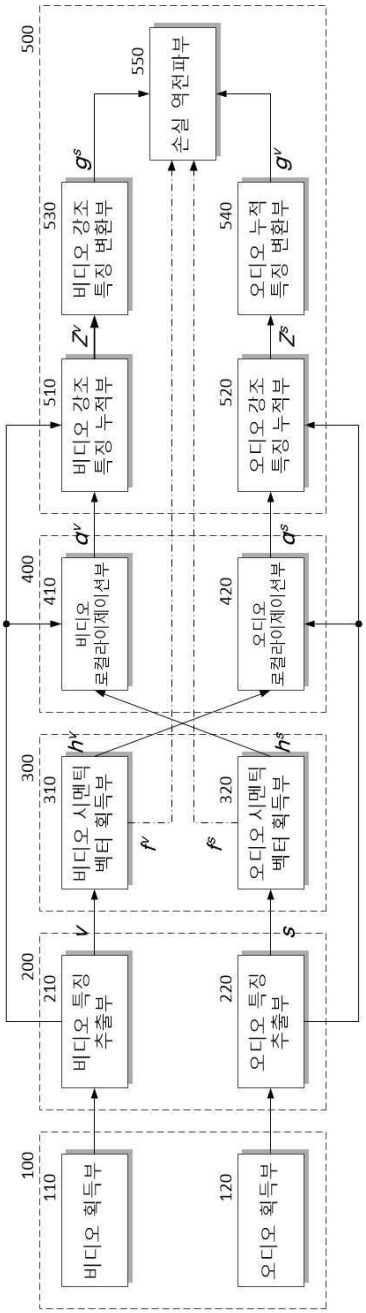
500: 학습부 310: 비디오 시멘틱 벡터 획득부

320: 오디오 시멘틱 벡터 획득부 410: 비디오 로컬라이제이션부

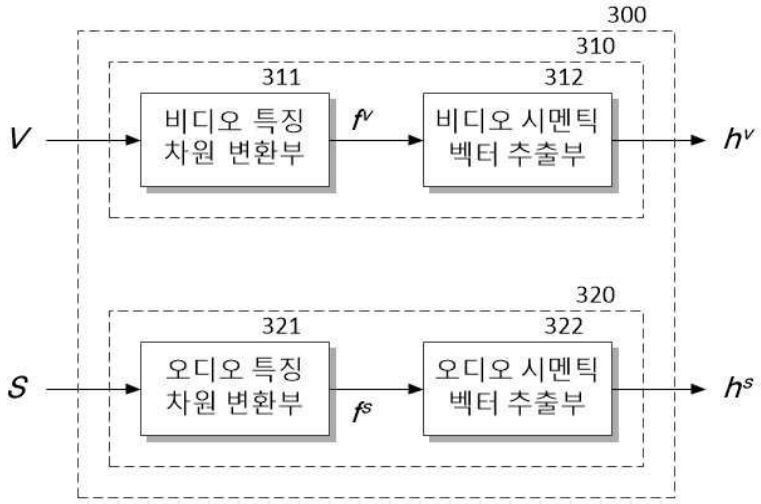
420: 오디오 로컬라이제이션부

도면

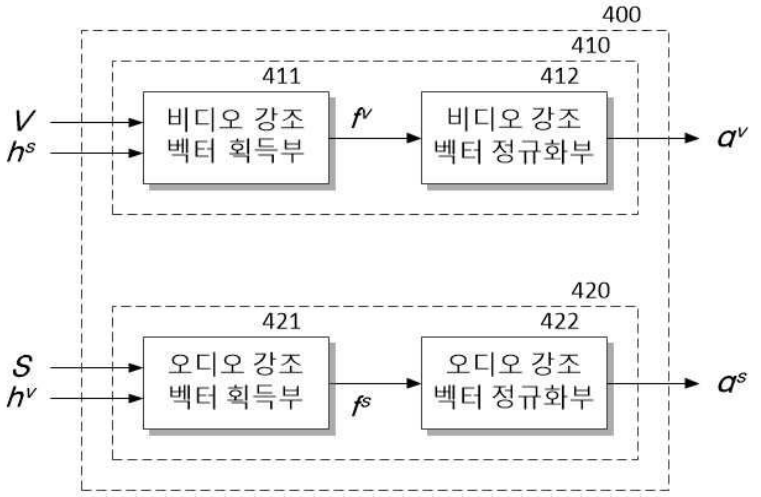
도면1



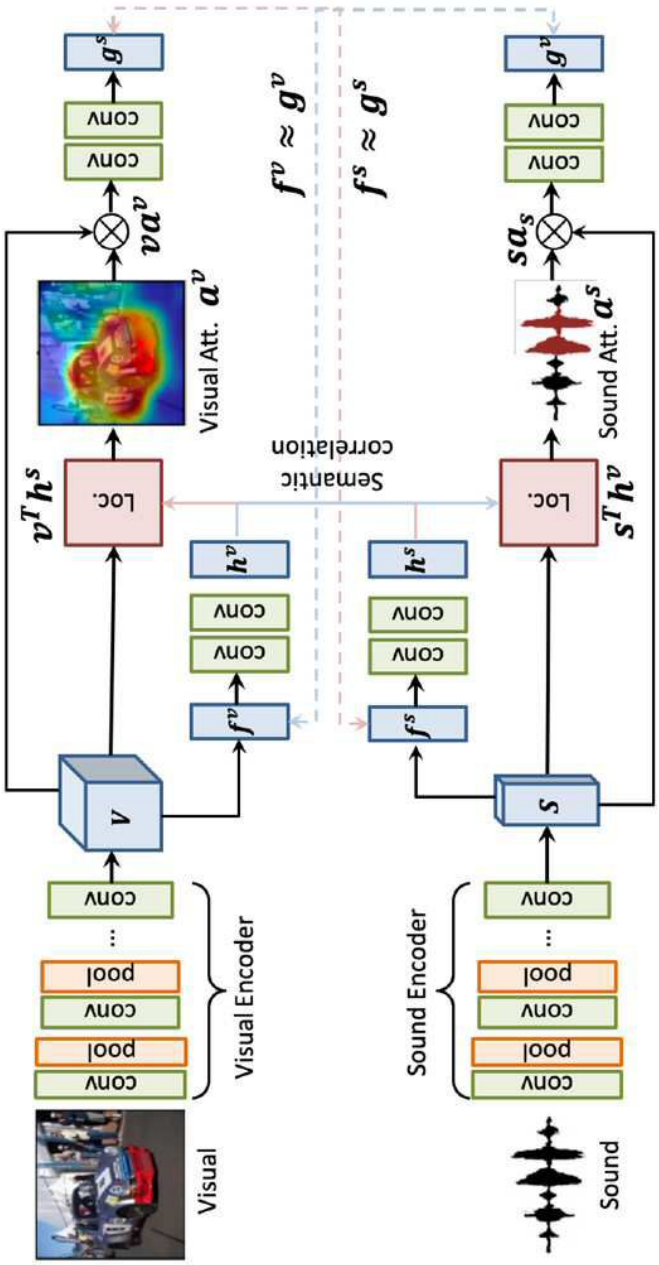
도면2


300
310
311
312
V
비디오 특징
차원 변환부
f^V
비디오 시멘틱
벡터 추출부
h^V
320
321
322
S
오디오 특징
차원 변환부
f^S
오디오 시멘틱
벡터 추출부
h^S

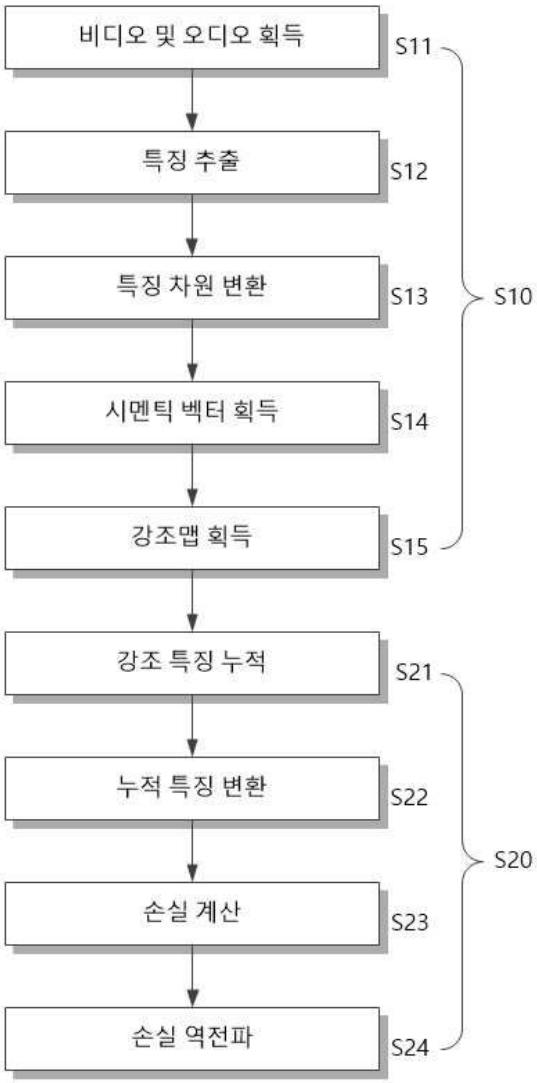
도면3


400
410
411
412
V
h^S
비디오 강조
벡터 획득부
f^V
비디오 강조
벡터 정규화부
a^V
420
421
422
S
h^V
오디오 강조
벡터 획득부
f^S
오디오 강조
벡터 정규화부
a^S

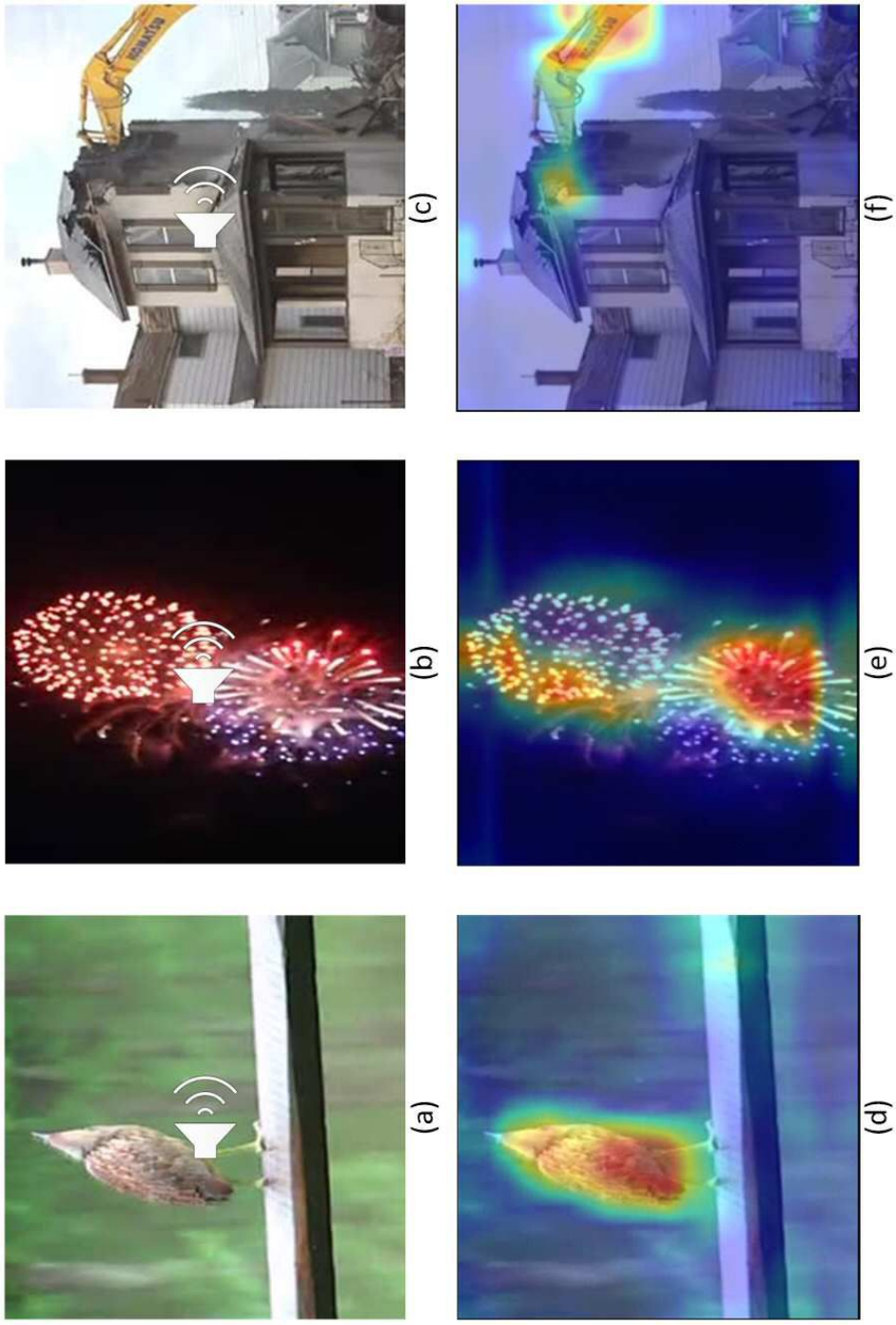
도면4


Visual
Visual Encoder
conv
pool
conv
pool
conv
...
conv
v
Visual Att.
Sound
Sound Encoder
conv
pool
conv
pool
conv
...
conv
s
Sound Att.
Loc.
Loc.
Semantic correlation
conv
conv
conv
conv

도면5


비디오 및 오디오 획득
S11
특징 추출
S12
특징 차원 변환
S13
S10
시멘틱 벡터 획득
S14
강조맵 획득
S15
강조 특징 누적
S21
누적 특징 변환
S22
S20
손실 계산
S23
손실 역전파
S24

도면6



도면7