(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년06월22일
(11) 등록번호 10-2546600
(24) 등록일자 2023년06월19일

(51) 국제특허분류(Int. Cl.)
G06F 18/00 (2023.01) *G06N 3/08* (2023.01)
G06T 15/08 (2011.01) *G06T 7/593* (2017.01)
G06V 10/40 (2022.01)
(52) CPC특허분류
G06V 40/174 (2022.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2020-0180996
(22) 출원일자 2020년12월22일
심사청구일자 2020년12월22일
(65) 공개번호 10-2022-0076247
(43) 공개일자 2022년06월08일
(30) 우선권주장
1020200164401 2020년11월30일 대한민국(KR)
(56) 선행기술조사문헌
한국 공개특허공보 제10-2019-0128933호(2019.11.19.) 1부.*
한국 등록특허공보 제10-2086067호(2020.03.06.) 1부.*
*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 18 항 심사관 : 정수진

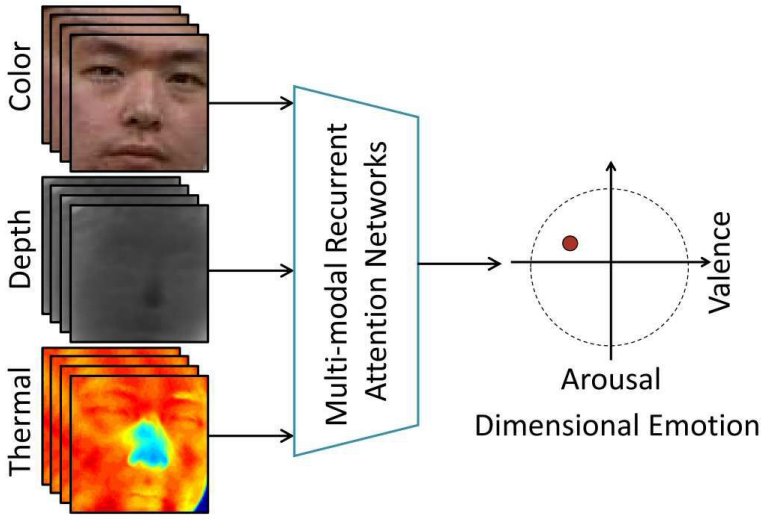
(54) 발명의 명칭 **멀티모달 융합 기반 감정 인식 장치 및 방법**

(57) 요 약

본 발명은 미리 학습된 인공 신경망으로 구현되어 각각 T개의 연속하는 프레임을 포함하는 깊이 시퀀스 영상과 열 시퀀스 영상 및 컬러 시퀀스 영상 각각을 프레임 순서에 따라 순차적으로 인가받아, 학습된 방식에 따라 인가되는 프레임들의 공간적 특징을 순차적으로 추출하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 획

(뒷면에 계속)

대 표 도 - 도1



득하는 공간적 인코더, 미리 학습된 인공 신경망으로 구현되어 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 순차적으로 인가받아, 학습된 방식에 따라 순차적으로 인가되는 특징맵들 사이의 시간적 특징을 추가하고 융합 디코딩하여 융합 컬러 시공간 특징을 획득하는 시간적 디코더, 미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 컬러 시퀀스 영상으로부터 3D 특징 볼륨을 추출하고, 순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여 시공간 주의 볼륨을 획득하며, 3D 특징 볼륨과 시공간 주의 볼륨을 결합하여 주의 강화 특징 볼륨을 획득하는 시공간 주의 볼륨 획득부 및 미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 시공간 주의 볼륨으로부터 감정값을 추정하여 획득하는 감정 추정부를 포함하여, 멀티모달 영상을 융합하여 감정을 인식하고 시간적 변화가 함께 반영되도록 하여 매우 정확하게 감정을 인식할 수 있는 감정 인식 장치 및 방법을 제공할 수 있다.

(52) CPC특허분류
G06T 15/08 (2013.01)
G06T 7/593 (2017.01)
G06V 10/40 (2023.01)
G06V 20/64 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116308
과제번호	2017M3C4A7069370
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	원천기술개발사업
연구과제명	(2세부)딥러닝 기반 의미론적 상황 이해 원천기술 연구 (2단계)(2/2)
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2020.04.01 ~ 2020.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

미리 학습된 인공 신경망으로 구현되어 각각 T개의 연속하는 프레임을 포함하는 깊이 시퀀스 영상과 열 시퀀스 영상 및 컬러 시퀀스 영상 각각을 프레임 순서에 따라 순차적으로 인가받아, 학습된 방식에 따라 인가되는 프레임들의 공간적 특징을 순차적으로 추출하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 획득하는 공간적 인코더;

미리 학습된 인공 신경망으로 구현되어 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 순차적으로 인가받아, 학습된 방식에 따라 순차적으로 인가되는 특징맵들 사이의 시간적 특징을 추가하고 융합 디코딩하여 융합 컬러 시공간 특징을 획득하는 시간적 디코더;

미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 컬러 시퀀스 영상으로부터 3D 특징 볼륨을 추출하고, 순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여 시공간 주의 볼륨을 획득하며, 3D 특징 볼륨과 시공간 주의 볼륨을 결합하여 주의 강화 특징 볼륨을 획득하는 시공간 주의 볼륨 획득부; 및

미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 시공간 주의 볼륨으로부터 감정값을 추정하여 획득하는 감정 추정부를 포함하되,

상기 시간적 디코더는

학습된 방식에 따라 순차적으로 인가되는 T개의 깊이 특징맵 각각에 대해 이전 인가된 깊이 특징맵에서 추출된 히든 특징을 이전 깊이 시공간 특징으로서 함께 디코딩하여 순차적으로 T개의 깊이 시공간 특징을 획득하는 깊이 디코더;

학습된 방식에 따라 순차적으로 인가되는 T개의 열 특징맵 각각에 대해 이전 인가된 열 특징맵에서 추출된 히든 특징을 이전 열 시공간 특징으로서 함께 디코딩하여 순차적으로 T개의 열 시공간 특징을 획득하는 열 디코더; 및

학습된 방식에 따라 순차적으로 인가되는 T개의 컬러 특징맵 각각과 대응하는 깊이 시공간 특징과 열 시공간 특징 및 이전 획득된 융합 컬러 시공간 특징이 융합된 융합 히든 특징을 함께 디코딩하여 순차적으로 T개의 융합 컬러 시공간 특징을 획득하는 융합 컬러 디코더를 포함하는 감정 인식 장치.

청구항 2

제1항에 있어서, 상기 공간적 인코더는

학습된 방식에 따라 상기 깊이 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 깊이 특징맵을 추출하는 깊이 인코더;

학습된 방식에 따라 상기 열 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 열 특징맵을 추출하는 열 인코더; 및

학습된 방식에 따라 상기 컬러 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 컬러 특징맵을 추출하는 컬러 인코더를 포함하는 감정 인식 장치.

청구항 3

삭제

청구항 4

제1항에 있어서, 상기 융합 컬러 디코더는

대응하는 깊이 시공간 특징과 열 시공간 특징 및 이전 획득된 융합 컬러 시공간 특징 각각에 대해 기지정된 가

중치로 가중하고 합하여 상기 융합 히든 특징을 획득하는 감정 인식 장치.

청구항 5

제1항에 있어서, 상기 깊이 디코더와 상기 열 디코더 및 상기 융합 컬러 디코더는 각각 ConvLSTM(Convolutional Long Short-Term Memory)으로 구현되는 감정 인식 장치.

청구항 6

제1항에 있어서, 상기 시공간 주의 볼륨 획득부는

미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 T개의 프레임을 포함하는 컬러 시퀀스 영상을 3D의 단일 이미지로 인지하여 특징을 추출하여 3D 특징 볼륨을 획득하는 3D 특징 추출부;

순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여, 시공간 특징 볼륨을 획득하고, 획득된 시공간 특징 볼륨을 기지정된 정규화하여 상기 시공간 주의 볼륨을 획득하는 정규화부; 및

상기 3D 특징 볼륨과 상기 시공간 주의 볼륨을 하다마드 곱셈하여 상기 주의 강화 특징 볼륨을 획득하는 주의 강화부를 포함하는 감정 인식 장치.

청구항 7

제6항에 있어서, 상기 3D 특징 추출부는

미리 학습된 3D CNN(3D Convolutional Neural Networks)으로 구현되는 감정 인식 장치.

청구항 8

제6항에 있어서, 상기 정규화부는

시공간 특징 볼륨(H)을 소프트 맥스 함수를 이용하여 수학식

$$A_{t,i} = \frac{\exp(H_{t,i})}{\sum_j \exp(H_{t,j})}$$

(여기서 $H_{t,i}$ 는 시공간 특징 볼륨(H)에 포함된 시간(t)에서의 융합 컬러 시공간 특징(h_t^I)인 시공간 특징맵(H_t)에서 i(i ∈ {1,ㅇㅇㅇ , H ㅅ W}) 픽셀 위치의 특징을 나타내고, $A_{t,i}$ 는 시공간 특징 볼륨(H)의 위치별 특징($H_{t,i}$)에 대응하는 시공간 주의 볼륨(A)의 가중치를 나타낸다.)

에 따라 정규화하는 감정 인식 장치.

청구항 9

제6항에 있어서, 상기 감정 추정부는

상기 시공간 주의 볼륨으로부터 각성(Arousal) 및 유인가(Valence)를 2개의 축으로 하는 2차원 상의 기지정된 범위 이내의 스칼라 좌표값으로 상기 감정값을 추정하여 획득하는 감정 인식 장치.

청구항 10

제1항에 있어서, 상기 감정 인식 장치는

상기 감정값(y)과 미리 획득된 진리값($\hat{y}$)을 비교하여 수학식

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M} \|\hat{y}_m - y_m\|_2$$

(여기서 $\| \|_2$ 는 L2-norm 함수이다.)

에 따라 손실(L)을 계산하고, 계산된 손실(L)을 역전파하여 학습을 수행하는 학습부를 더 포함하는 감정 인식 장치.

청구항 11

미리 학습된 인공 신경망을 이용하여 각각 T개의 연속하는 프레임을 포함하는 깊이 시퀀스 영상과 열 시퀀스 영상 및 컬러 시퀀스 영상 각각을 프레임 순서에 따라 순차적으로 인가받아, 학습된 방식에 따라 인가되는 프레임들의 공간적 특징을 순차적으로 추출하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 획득하는 단계;

미리 학습된 인공 신경망을 이용하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 순차적으로 인가받아, 학습된 방식에 따라 순차적으로 인가되는 특징맵들 사이의 시간적 특징을 추가하고 융합 디코딩하여 융합 컬러 시공간 특징을 획득하는 단계;

미리 학습된 인공 신경망을 이용하여 학습되는 방식에 따라 컬러 시퀀스 영상으로부터 3D 특징 볼륨을 추출하고, 순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여 시공간 주의 볼륨을 획득하며, 3D 특징 볼륨과 시공간 주의 볼륨을 결합하여 주의 강화 특징 볼륨을 획득하는 단계; 및

미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 시공간 주의 볼륨으로부터 감정값을 추정하여 획득하는 단계를 포함하되,

상기 융합 컬러 시공간 특징을 획득하는 단계는

학습된 방식에 따라 순차적으로 인가되는 T개의 깊이 특징맵 각각에 대해 이전 인가된 깊이 특징맵에서 추출된 히든 특징을 이전 깊이 시공간 특징으로서 함께 디코딩하여 순차적으로 T개의 깊이 시공간 특징을 획득하는 단계;

학습된 방식에 따라 순차적으로 인가되는 T개의 열 특징맵 각각에 대해 이전 인가된 열 특징맵에서 추출된 히든 특징을 이전 열 시공간 특징으로서 함께 디코딩하여 순차적으로 T개의 열 시공간 특징을 획득하는 단계; 및

학습된 방식에 따라 순차적으로 인가되는 T개의 컬러 특징맵 각각과 대응하는 깊이 시공간 특징과 열 시공간 특징 및 이전 획득된 융합 컬러 시공간 특징이 융합된 융합 히든 특징을 함께 디코딩하여 순차적으로 T개의 융합 컬러 시공간 특징을 획득하는 단계를 포함하는 감정 인식 방법.

청구항 12

제11항에 있어서, 상기 컬러 특징맵을 획득하는 단계는

학습된 방식에 따라 상기 깊이 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 깊이 특징맵을 추출하는 단계;

학습된 방식에 따라 상기 열 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 열 특징맵을 추출하는 단계; 및

학습된 방식에 따라 상기 컬러 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 컬러 특징맵을 추출하는 단계를 포함하는 감정 인식 방법.

청구항 13

삭제

청구항 14

제11항에 있어서, 상기 융합 히든 특징은

대응하는 깊이 시공간 특징과 열 시공간 특징 및 이전 획득된 융합 컬러 시공간 특징 각각에 대해 기지정된 가중치로 가중하고 합하여 획득되는 감정 인식 방법.

청구항 15

제11항에 있어서, 상기 융합 컬러 시공간 특징을 획득하는 단계는

다수의 ConvLSTM(Convolutional Long Short-Term Memory)을 이용하여, 상기 T개의 깊이 시공간 특징과 상기 T개의 열 시공간 특징 및 상기 T개의 융합 컬러 시공간 특징을 획득하는 감정 인식 방법.

청구항 16

제11항에 있어서, 상기 주의 강화 특징 볼륨을 획득하는 단계는

미리 학습된 인공 신경망을 이용하여 학습되는 방식에 따라 T개의 프레임을 포함하는 컬러 시퀀스 영상을 3D의 단일 이미지로 인지하여 특징을 추출하여 3D 특징 볼륨을 획득하는 단계;

순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여, 시공간 특징 볼륨을 획득하고, 획득된 시공간 특징 볼륨을 기지정된 정규화하여 상기 시공간 주의 볼륨을 획득하는 단계; 및

상기 3D 특징 볼륨과 상기 시공간 주의 볼륨을 하다마드 곱셈하여 상기 주의 강화 특징 볼륨을 획득하는 단계를 포함하는 감정 인식 방법.

청구항 17

제16항에 있어서, 상기 3D 특징 볼륨을 획득하는 단계는

미리 학습된 3D CNN(3D Convolutional Neural Networks)을 이용하여 상기 3D 특징 볼륨을 획득하는 감정 인식 방법.

청구항 18

제16항에 있어서, 상기 시공간 주의 볼륨을 획득하는 단계는

시공간 특징 볼륨(H)을 소프트 맥스 함수를 이용하여 수학식

$$A_{t,i} = \frac{\exp(H_{t,i})}{\sum_j \exp(H_{t,j})}$$

(여기서 $H_{t,i}$ 는 시공간 특징 볼륨(H)에 포함된 시간(t)에서의 융합 컬러 시공간 특징(h_t^I)인 시공간 특징맵(H_t)에서 i(i ∈ {1,ㅇㅇㅇ , H ㅅ W}) 픽셀 위치의 특징을 나타내고, $A_{t,i}$ 는 시공간 특징 볼륨(H)의 위치별 특징($H_{t,i}$)에 대응하는 시공간 주의 볼륨(A)의 가중치를 나타낸다.)

에 따라 정규화하는 감정 인식 방법.

청구항 19

제16항에 있어서, 상기 감정값을 추정하여 획득하는 단계는

상기 시공간 주의 볼륨으로부터 각성(Arousal) 및 유인가(Valence)를 2개의 축으로 하는 2차원 상의 기지정된 범위 이내의 스칼라 좌표값으로 상기 감정값을 추정하여 획득하는 감정 인식 방법.

청구항 20

제11항에 있어서, 상기 감정 인식 방법은

상기 감정값(y)과 미리 획득된 진리값($\hat{y}$)을 비교하여 수학식

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M} \|\hat{y}_m - y_m\|_2$$

(여기서 $\| \|_2$ 는 L2-norm 함수이다.)

에 따라 손실(L)을 계산하고, 계산된 손실(L)을 역전파하여 학습을 수행하는 단계를 더 포함하는 감정 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 감정 인식 장치 및 방법에 관한 것으로, 멀티모달 융합 기반 감정 인식 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 인간의 감정은 대인관계에서 중요한 역할을 한다. 감정은 사람의 얼굴, 행동, 말과 같은 요소로 표현될 수 있으며, 향후 로봇-사람 과의 상호소통에서도 중요한 역할을 하게 될 것으로 예측되고 있다. 특히 시각적 콘텐츠에서 인간의 감정을 이해하는 것은 건강, 개인 보조, 정서적 컴퓨팅 및 이미지 처리 분야 등에서 다양하게 이용될 수 있다.

[0003] 현재 감정인식 기법은 크게 클래스화 감정인식(Categorical emotion recognition)과 연속적인 감정 인식(Dimensional emotion recognition)으로 구분될 수 있다. 클래스화 감정 인식은 감정을 공포, 분노, 행복, 혐오, 슬픔, 놀람과 같은 6가지 기본 감정에 따라 이산된 범주로 구분한다. 다만 클래스화 감정 인식은 지정된 감정으로만 분류하여 인식함에 따라 분류되지 않는 감정의 영역이 존재될 뿐만 아니라, 인식 가능한 감정의 종류가 제한되는 한계가 있다.

[0004] 그에 반해 연속적인 감정 인식의 경우, 긍정적/부정적 감정을 나타내는 발란스(valence)와 진정/흥분 정도를 나타내는 흥분도(arusal)를 측정하여 사람의 감정 정도를 나타내도록 한다. 여기서 발란스와 흥분도의 두 값은 모두 -1과 1 사이의 연속적인 값을 가질 수 있다. 즉 연속적인 감정 인식은 인간의 감정을 발란스와 흥분도의 2차원 공간 상에서 제한없이 표현이 가능함에 따라 다양한 감정을 표현을 인식할 수 있다는 장점이 있다. 다만 클래스화 감정인식과 연속적인 감정 인식 모두 감정을 공간적 정보로서 효과적으로 추출하도록 구상된 반면, 표정 요인의 시간적 변화를 정확하게 반영하지 못하는 한계가 있다.

[0005] 한편, 기존의 감정 인식은 주로 사람 얼굴에 대한 컬러 영상에 기반하여 감정을 인식하도록 학습된 인공 신경망을 이용하여 수행되고 있다. 그러나 컬러 영상만을 입력으로 받는 인공 신경망을 이용하여 감정 인식을 수행하는 경우, 대상자의 피부색, 조도, 조명 등에 의한 영향을 크게 받으며, 이로 인해 감정을 오인식하게 되는 경우가 빈번하게 발생하는 문제가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 등록 특허 제10-2060719호 (2019.12.23 등록)

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은 컬러 영상뿐만 아니라 깊이 영상과 열 영상이 함께 포함된 멀티모달 영상을 융합하여 정확하게 감정을 인식할 수 있는 감정 인식 장치 및 방법을 제공하는데 있다.

[0008] 본 발명의 다른 목적은 시간적 변화를 반영하여 더욱 정확하게 감정을 인식할 수 있는 감정 인식 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 감정 인식 장치는 미리 학습된 인공 신경망으로 구현되어 각각 T개의 연속하는 프레임을 포함하는 깊이 시퀀스 영상과 열 시퀀스 영상 및 컬러 시퀀스 영상 각각을 프

레임 순서에 따라 순차적으로 인가받아, 학습된 방식에 따라 인가되는 프레임들의 공간적 특징을 순차적으로 추출하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 획득하는 공간적 인코더; 미리 학습된 인공 신경망으로 구현되어 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 순차적으로 인가받아, 학습된 방식에 따라 순차적으로 인가되는 특징맵들 사이의 시간적 특징을 추가하고 융합 디코딩하여 융합 컬러 시공간 특징을 획득하는 시간적 디코더; 미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 컬러 시퀀스 영상으로부터 3D 특징 볼륨을 추출하고, 순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여 시공간 주의 볼륨을 획득하며, 3D 특징 볼륨과 시공간 주의 볼륨을 결합하여 주의 강화 특징 볼륨을 획득하는 시공간 주의 볼륨 획득부; 및 미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 시공간 주의 볼륨으로부터 감정값을 추정하여 획득하는 감정 추정부를 포함한다.

[0010] 상기 공간적 인코더는 학습된 방식에 따라 상기 깊이 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 깊이 특징맵을 추출하는 깊이 인코더; 학습된 방식에 따라 상기 열 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 열 특징맵을 추출하는 열 인코더; 및 학습된 방식에 따라 상기 컬러 시퀀스 영상의 T개의 프레임 각각에 대한 공간적 특징을 순차적으로 추출하여 T개의 컬러 특징맵을 추출하는 컬러 인코더를 포함할 수 있다.

[0011] 상기 시간적 디코더는 학습된 방식에 따라 순차적으로 인가되는 T개의 깊이 특징맵 각각에 대해 이전 인가된 깊이 특징맵에서 추출된 히든 특징을 이전 깊이 시공간 특징으로서 함께 디코딩하여 순차적으로 T개의 깊이 시공간 특징을 획득하는 깊이 디코더; 학습된 방식에 따라 순차적으로 인가되는 T개의 열 특징맵 각각에 대해 이전 인가된 열 특징맵에서 추출된 히든 특징을 이전 열 시공간 특징으로서 함께 디코딩하여 순차적으로 T개의 열 시공간 특징을 획득하는 열 디코더; 및 학습된 방식에 따라 순차적으로 인가되는 T개의 컬러 특징맵 각각과 대응하는 깊이 시공간 특징과 열 시공간 특징 및 이전 획득된 융합 컬러 시공간 특징이 융합된 융합 히든 특징을 함께 디코딩하여 순차적으로 T개의 융합 컬러 시공간 특징을 획득하는 융합 컬러 디코더를 포함할 수 있다.

[0012] 상기 융합 컬러 디코더는 대응하는 깊이 시공간 특징과 열 시공간 특징 및 이전 획득된 융합 컬러 시공간 특징 각각에 대해 기지정된 가중치로 가중하고 합하여 상기 융합 히든 특징을 획득할 수 있다.

[0013] 상기 깊이 디코더와 상기 열 디코더 및 상기 융합 컬러 디코더는 각각 ConvLSTM(Convolutional Long Short-Term Memory)으로 구현될 수 있다.

[0014] 상기 시공간 주의 볼륨 획득부는 미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 T개의 프레임을 포함하는 컬러 시퀀스 영상을 3D의 단일 이미지로 인지하여 특징을 추출하여 3D 특징 볼륨을 획득하는 3D 특징 추출부; 순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여, 시공간 특징 볼륨을 획득하고, 획득된 시공간 특징 볼륨을 기지정된 정규화하여 상기 시공간 주의 볼륨을 획득하는 정규화부; 및 상기 3D 특징 볼륨과 상기 시공간 주의 볼륨을 하다마드 곱셈하여 상기 주의 강화 특징 볼륨을 획득하는 주의 강화부를 포함할 수 있다.

[0015] 상기 3D 특징 추출부는 미리 학습된 3D CNN(3D Convolutional Neural Networks)으로 구현될 수 있다.

[0016] 상기 감정 추정부는 상기 시공간 주의 볼륨으로부터 각성(Arousal) 및 유인가(Valence)를 2개의 축으로 하는 2차원 상의 기지정된 범위 이내의 스칼라 좌표값으로 상기 감정값을 추정하여 획득할 수 있다.

[0017] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 감정 인식 방법은 미리 학습된 인공 신경망을 이용하여 각각 T개의 연속하는 프레임을 포함하는 깊이 시퀀스 영상과 열 시퀀스 영상 및 컬러 시퀀스 영상 각각을 프레임 순서에 따라 순차적으로 인가받아, 학습된 방식에 따라 인가되는 프레임들의 공간적 특징을 순차적으로 추출하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 획득하는 단계; 미리 학습된 인공 신경망을 이용하여 각각 T개의 깊이 특징맵과 열 특징맵 및 컬러 특징맵을 순차적으로 인가받아, 학습된 방식에 따라 순차적으로 인가되는 특징맵들 사이의 시간적 특징을 추가하고 융합 디코딩하여 융합 컬러 시공간 특징을 획득하는 단계; 미리 학습된 인공 신경망을 이용하여 학습되는 방식에 따라 컬러 시퀀스 영상으로부터 3D 특징 볼륨을 추출하고, 순차적으로 획득되는 T개의 융합 컬러 시공간 특징을 누적하여 시공간 주의 볼륨을 획득하며, 3D 특징 볼륨과 시공간 주의 볼륨을 결합하여 주의 강화 특징 볼륨을 획득하는 단계; 및 미리 학습된 인공 신경망으로 구현되어 학습되는 방식에 따라 시공간 주의 볼륨으로부터 감정값을 추정하여 획득하는 단계를 포함한다.

발명의 효과

[0018] 따라서, 본 발명의 실시예에 따른 감정 인식 장치 및 방법은 컬러 영상과 깊이 영상 및 열 영상을 포함하는 멀

티모달 영상을 인가받고, 멀티모달 영상을 융합하여 감정을 인식함으로써 매우 정확하게 감정을 인식할 수 있다. 뿐만 아니라, 영상에 포함된 다수 프레임에 대한 시간적 변화가 함께 반영되도록 하여 감정 인식 정확도를 더욱 향상시킬 수 있다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 일 실시예에 따른 감정 인식 장치의 동작 개념을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 감정 인식 장치의 개략적 구성을 나타낸다.

도 3은 도 2의 감정 인식 장치의 각 구성별 동작을 설명하기 위한 도면이다.

도 4는 도 2의 시간적 디코더의 상세 구성을 나타낸다.

도 5는 본 실시예의 감정 인식 장치에 입력되는 멀티모달 영상과 어텐션 스코어 맵의 일 예를 나타낸다.

도 6은 본 발명의 일 실시예에 따른 감정 인식 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0020] 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.

[0021] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러나, 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.

[0022] 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.

[0023] 도 1은 본 발명의 일 실시예에 따른 감정 인식 장치의 동작 개념을 설명하기 위한 도면이다.

[0024] 도 1을 참조하면, 본 실시예에 따른 감정 인식 장치(10)는 인공 신경망으로 구현되며, 컬러 영상뿐만 아니라 깊이 영상 및 열 영상을 함께 인가받아 영상에 포함된 대상자의 감정을 추정하며, 추정된 감정을 발란스와 흥분도로 구성되는 2차원 공간 상에 특정 포인트로서 인식된 감정을 표현한다.

[0025] 기존의 감정 인식 장치(10) 또한 인공 신경망으로 구현되는 경우가 많았으나, 기존의 감정 인식 장치(10)는 주로 컬러 영상만을 인가받고, 인가된 컬러 영상으로부터 대상자의 감정을 추정하였다. 그러나 상기한 바와 같이, 컬러 영상만으로 대상자의 감정을 추정하는 경우, 피부색, 조도, 조명과 같은 환경 요인에 의한 변화가 크기 때문에 대상자의 감정을 잘못 인식할 수 있다.

[0026] 이에 본 실시예에 따른 감정 인식 장치(10)는 컬러 영상과 함께 깊이 영상 및 열 영상을 더 인가받고, 인가된 깊이 영상 및 열 영상을 기반으로 컬러 영상에서 주의해야할 영역을 가이드하도록 함으로써, 대상자의 감정을 정확하게 인식할 수 있도록 한다.

[0027] 또한 본 실시예에 따른 감정 인식 장치(10)는 컬러 영상과 깊이 영상 및 열 영상 각각에 대한 단일 프레임 영상을 인가받는 것이 아니라, 다수의 프레임으로 구성되는 시퀀스 영상으로 인가받고, 시퀀스 영상으로부터 시간적 특징이 반영되도록 이전 프레임의 특성을 다음 프레임에 순환 적용하여 특징을 추출함으로써, 대상자의 감정을 더욱 정확하게 인식할 수 있도록 한다.

[0028] 도 2는 본 발명의 일 실시예에 따른 감정 인식 장치의 개략적 구성을 나타내고, 도 3은 도 2의 감정 인식 장치의 각 구성별 동작을 설명하기 위한 도면이며, 도 4는 도 2의 시간적 디코더의 상세 구성을 나타낸다. 그리고 도 5는 본 실시예의 감정 인식 장치에 입력되는 멀티모달 영상과 어텐션 스코어 맵의 일 예를 나타낸다.

[0029] 도 2를 참조하면, 본 실시예에 따른 감정 인식 장치(10)는 멀티모달 영상 획득부(100), 공간적 인코더(200), 시간적 디코더(300), 시공간 주의 볼륨 획득부(400) 및 감정 추정부(500)를 포함할 수 있다.

[0030] 우선 멀티모달 영상 획득부(100)는 감정 인식 대상의 얼굴이 포함된 멀티모달 영상을 획득한다. 여기서 멀티모

달 영상 획득부(100)는 컬러 영상 획득부(130)와 깊이 영상 획득부(110) 및 열 영상 획득부(120)를 포함할 수 있다. 컬러 영상 획득부(130)는 도 5의 (a)에 도시된 바와 같이, 대상자의 얼굴이 촬영된 컬러 영상(I)을 획득한다. 그리고 깊이 영상 획득부(110)와 열 영상 획득부(120) 각각은 도 5의 (b)와 (c)에 도시된 바와 같이 컬러 영상 획득부(130)에서 획득되는 컬러 영상(I)과 동일한 대상자에 대한 깊이 영상(D)과 열 영상(F)을 획득한다.

[0031] 즉 멀티모달 영상 획득부(100)는 컬러 영상뿐만 아니라, 깊이 영상과 열 영상이 함께 포함되는 멀티모달 영상을 획득한다.

[0032] 여기서 컬러 영상과 깊이 영상 및 열 영상은 도 5의 (a) 내지 (c)에 도시된 바와 같이, 동일한 대상자에 대해 기지정된 동일한 기간 동안 촬영되어 획득되는 시퀀스 영상으로 각각 연속하는 T개의 프레임을 갖는 시퀀스 영상일 수 있다. 즉 컬러 영상과 깊이 영상 및 열 영상 각각은 T개의 프레임을 포함하는 컬러 시퀀스 영상($I_{1:T}$ = {I_1, I_2, …, I_T})과 깊이 시퀀스 영상($D_{1:T}$ = {D_1, D_2, …, D_T}) 및 열 시퀀스 영상($F_{1:T}$ = {F_1, F_2, …, F_T})으로 획득된다.

[0033] 멀티모달 영상 획득부(100)는 컬러 시퀀스 영상과 깊이 시퀀스 영상 및 열 시퀀스 영상 각각에서 T개를 초과하는 프레임으로 획득할 수 있으나, 획득된 컬러 시퀀스 영상과 깊이 시퀀스 영상 및 열 시퀀스 영상 각각에 포함된 프레임의 개수가 T개를 초과하는 경우, 멀티모달 영상 획득부(100)는 대상의 감정을 인식하고자 하는 시점의 프레임이 포함된 T개의 프레임을 추출하여 인코더(200)로 전달할 수 있다.

[0034] 공간적 인코더(200)는 멀티모달 영상 획득부(100)에서 획득된 멀티모달 영상을 프레임 단위로 인가받고, 인가된 멀티모달 영상을 미리 학습된 방식으로 인코딩하여 2차원 공간 특징을 추출한다. 공간적 인코더(200)는 각각 미리 학습된 인공 신경망으로 구현되는 깊이 인코더(210), 열 인코더(220) 및 컬러 인코더(230)를 포함하여, 멀티모달 영상에 포함된 깊이 영상과 열 영상 및 컬러 영상 각각을 미리 학습된 방식에 따라 인코딩함으로써 깊이 특징맵(x^D), 열 특징맵(x^F) 및 컬러 특징맵(x^I)을 획득한다.

[0035] 깊이 인코더(210)는 T개의 프레임을 갖는 깊이 시퀀스 영상($D_{1:T}$ = {D_1, D_2, …, D_T})에서 각 프레임에 해당하는 깊이 영상(D_1, D_2, …, D_T)을 순차적으로 인가받고, 순차적으로 인가되는 깊이 영상(D_1, D_2, …, D_T) 각각에 대해 미리 학습된 방식에 따라 인코딩하여 각 깊이 영상의 공간적 특징을 추출함으로써 깊이 특징맵(x^D)을 획득한다.

[0036] 열 인코더(220)는 T개의 프레임을 갖는 열 시퀀스 영상($F_{1:T}$ = {F_1, F_2, …, F_T})에서 각 프레임에 해당하는 열 영상(F_1, F_2, …, F_T)을 순차적으로 인가받고, 순차적으로 인가되는 열 영상(F_1, F_2, …, F_T) 각각에 대해 미리 학습된 방식에 따라 인코딩하여 각 열 영상의 공간적 특징을 나타내는 열 특징맵(x^F)을 획득한다.

[0037] 마찬가지로 컬러 인코더(230) 또한 T개의 프레임을 갖는 컬러 시퀀스 영상($I_{1:T}$ = {I_1, I_2, …, I_T})의 각 컬러 영상을 순차적으로 인가받아 미리 학습된 방식에 따라 인코딩하여 각 컬러 영상의 공간적 특징을 나타내는 컬러 특징맵(x^I)을 획득한다.

[0038] 여기서 획득되는 깊이 특징맵(x^D)과 열 특징맵(x^F) 및 컬러 특징맵(x^I)은 2차원 특징맵의 형태로 획득될 수 있다.

[0039] 즉 공간적 인코더(200)는 깊이 시퀀스 영상($D_{1:T}$)과 열 시퀀스 영상($F_{1:T}$) 및 컬러 시퀀스 영상($I_{1:T}$ = {I_1, I_2, …, I_T})의 각 프레임에 대해 공간적 특징을 추출하여, 각각 T개의 깊이 특징맵(x^D)과 열 특징맵(x^F) 및 컬러 특징맵(x^I)을 획득한다. 여기서 깊이 특징맵(x^D)과 열 특징맵(x^F) 및 컬러 특징맵(x^I)을 통합하여 멀티모달 특징맵이라 할 수 있다. 따라서 공간적 인코더(200)는 T개의 프레임을 포함하는 멀티모달 영상(D, F, I)을 인가받아 T개의 멀티모달 특징맵(x^D, x^F, x^I)을 획득할 수 있다.

[0040] 공간적 인코더(200)의 깊이 인코더(210), 열 인코더(220) 및 컬러 인코더(230) 각각은 일 예로 미리 학습된 2차원 콘볼루션 신경망(2D Convolutional Neural Networks: 이하 2D CNN)으로 구현될 수 있으며, 맥스 풀링 레이어(max-pooling layer)를 더 포함할 수 있다.

[0041] 시간적 디코더(300)는 공간적 인코더(200)에서 획득된 T개의 멀티모달 특징맵(x^D, x^F, x^I)을 인가받고, T개의 멀티모달 특징맵(x^D, x^F, x^I)에 시간적 특징을 더 반영하여 융합함으로써, 시공간 주의맵을 획득한다.

[0042] 시간적 디코더(300) 또한 공간적 인코더(200)와 유사하게 각각 미리 학습된 인공 신경망으로 구현되는 깊이 디코더(310), 열 디코더(320) 및 융합 컬러 디코더(330)를 포함한다.

[0043] 우선 깊이 디코더(310)는 T개의 깊이 특징맵(x^D)을 순차적으로 인가받아 미리 학습된 방식에 따라 시간적 특성이 반영되도록 디코딩하여 깊이 시공간 특징(h^F)을 획득하고, 열 디코더(320)는 T개의 열 특징맵(x^F)을 순차적으로 인가받아 미리 학습된 방식에 따라 시간적 특성이 반영되도록 디코딩하여 열 시공간 특징(h^F)을 획득한다.

[0044] 다만 융합 컬러 디코더(330)는 깊이 디코더(310)나 열 디코더(320)와 달리 T개의 멀티모달 특징맵(x^D, x^F, x^I)에서 컬러 특징맵(x^I)뿐만 아니라, 깊이 디코더(310)에서 획득된 깊이 시공간 특징(h^F)과 열 시공간 특징(h^F)을 함께 순차적으로 인가받고, 인가된 컬러 특징맵(x^I)과 깊이 시공간 특징(h^F)과 열 시공간 특징(h^F)을 융합하고, 시간적 특성이 반영되도록 디코딩하여, 융합 컬러 시공간 특징(h^I)을 획득한다.

[0045] 여기서 깊이 디코더(310), 열 디코더(320)는 융합 컬러 디코더(330)가 컬러 특징맵(x^I)을 디코딩할 때, 컬러 영상(I)로부터 추정하기 어렵지만 깊이 영상(D)이나 열 영상(F)로부터 추정하기 용이한 시공간 특징 영역에 주의(attention)가 집중될 수 있도록 가이드하기 위한 깊이 시공간 특징(h^F)과 열 시공간 특징(h^F)을 제공하기 위한 구성으로 볼 수 있다.

[0046] 그리고 깊이 디코더(310), 열 디코더(320) 및 융합 컬러 디코더(330) 각각은 인공 신경망 중에서도 순차적으로 인가되는 T개의 멀티모달 특징맵(x^D, x^F, x^I)의 시간적 특징이 반영되도록 ConvLSTM(Convolutional Long Short-Term Memory)을 기반으로 구현될 수 있다. 여기서 ConvLSTM은 순환 신경망(Recurrent Neural Network: RNN)이 장기간(Long Term) 특징을 반영할 수 있도록 개선한 LSTM(Long Short-Term Memory)을 더욱 개선하여 공간적 특징을 더 반영할 수 있도록 구성되는 신경망이다. 시간적 디코더(300)의 깊이 디코더(310), 열 디코더(320) 및 컬러 디코더(330)가 시간적 특징을 반영할 수 있는 LSTM이 아닌 ConvLSTM을 기반으로 구현되는 것은 공간적 인코더(200)에서 획득된 T개의 멀티모달 특징맵(x^D, x^F, x^I)의 공간적 특징을 최대한 유지하면서 시간적 특징을 반영할 수 있도록 하기 위해서이다.

[0047] 도 4를 참조하면, 본 실시예에서 ConvLSTM을 기반으로 구현되는 깊이 디코더(310)와 열 디코더(320)를 주의 LSTM(Attention LSTM: 이하 A-LSTM)이라 하며, 각각 하나의 인공 신경망 셀(311, 321)만을 구비하여 이전 출력을 다시 인가받는 시간에 따른 순환 구조로 구성될 수 있으나, 경우에 따라서는 깊이 디코더(310)와 열 디코더(320) 각각은 다수(일예로 T개)의 인공 신경망 셀(311, 321)을 포함하도록 구성될 수도 있다. 그리고 도 4에 도시된 바와 같이, 깊이 디코더(310)와 열 디코더(320)는 동일한 구조의 인공 신경망 셀을 포함하도록 구성될 수 있다.

[0048] A-LSTM으로 구현되는 깊이 디코더(310)와 열 디코더(320)가 수행하는 기능은 수학식 1로 표현될 수 있다.

수학식 1

[0049]
$$
\begin{aligned}
i_t &= \sigma(w_{xi} * x_t + w_{hi} * h_{t-1} + w_{ci} * c_{t-1} + b_i),\\
f_t &= \sigma(w_{xf} * x_t + w_{hf} * h_{t-1} + w_{cf} * c_{t-1} + b_f),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(w_{xc} * x_t + w_{hc} * h_{t-1} + b_c),\\
o_t &= \sigma(w_{xo} * x_t + w_{ho} * h_{t-1} + w_{co} \odot c_t + b_o),\\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

[0050] 여기서 i_t, f_t, o_t, c_t 및 h_t 는 각각 시간 t에서 입력 게이트(input gate), 망각 게이트(forget gate), 출력 게

이트(output gate), 활성화 셀(activation cell) 및 셀 출력(cell output)을 나타낸다. 그리고 σ(·)와 tanh(·)는 각각 시그모이드(sigmoid) 함수와 쌍곡 탄젠트 함수(hyperbolic tangent)를 나타내며, *는 컨볼루션 연산자이고, ⊙는 하다마드(Hadamard) 곱셈 연산자를 나타낸다. 그리고 w은 다른 게이트를 연결하는 필터 행렬이고, b는 각 게이트에 상응하는 바이어스 벡터를 나타낸다.

[0051] 수학식 1과 도 4의 깊이 디코더(310)와 열 디코더(320)의 A-LSTM으로 구현되는 인공 신경망 셀을 살펴보면, 깊이 디코더(310)와 열 디코더(320)의 인공 신경망 셀은 각각 이전 시점(또는 프레임)(t-1) 셀의 출력 게이트(c_{t-1})와 셀 출력(h_{t-1}), 현재 시점(t)에서 대응하는 특징맵(x^D, x^F)을 인가받아, 수학식 1에 따른 연산을 수행하여, 현재 시점(t)의 출력 게이트(c_t)와 셀 출력(h_t)을 출력한다.

[0052] 여기서 셀 출력(h_t)은 A-LSTM의 히든 특징을 나타낸다. 즉 깊이 특징맵(x^D)과 열 특징맵(x^F)의 히든 특징(h_t^D, h_t^F)이다. 여기서는 깊이 특징맵(x^D)과 열 특징맵(x^F)이 공간적 인코더(200)에 의해 깊이 영상(D)과 열 영상(F)에서 공간적 특징이 추출된 특징맵이므로, 히든 특징(h_t^D, h_t^F)은 깊이 시공간 특징과 열 시공간 특징으로 볼 수 있다.

[0053] 한편, 융합 컬러 디코더(330)는 깊이 디코더(310)와 열 디코더(320)와 유사하게 ConvLSTM을 기반으로 구현되지만, 깊이 디코더(310)와 열 디코더(320)의 대응하는 인공 신경망 셀에서 획득되는 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F)을 컬러 특징맵(x^I)과 함께 인가받으며, 이전 획득된 융합 컬러 히든 특징(h_t^C)까지 함께 인가받는다는 점에서 일부 상이한 구성을 갖는다. 융합 컬러 히든 특징(h_t^C) 또한 융합 컬러 시공간 특징이라 할 수 있다.

[0054] 여기서 인가된 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F)은 가이드 특징으로서 인가되며, 융합 컬러 디코더(330)의 인공 신경망 셀에는 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F) 및 융합 컬러 시공간 특징(h_{t-1}^C) 각각은 기지정된 가중치로 가중되고 합쳐져서 융합 히든 특징($h_{t-1}^{I,D,F}$)으로 인가될 수 있다. 여기서 융합 컬러 시공간 특징(h_t^C)이라 하는 것은, 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F)이 컬러 히든 특징에 융합되어 반영되기 때문이다.

[0055] 도 5에서 (d)는 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F)이 융합되지 않은 컬러 시공간 특징을 나타내고, (e)는 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F)이 융합된 융합 컬러 시공간 특징(h_t^C)을 나타낸다. 도 5의 (d)와 (e)를 비교하면, 깊이 시공간 특징(h_t^D)과 열 시공간 특징(h_t^F)이 융합된 융합 컬러 시공간 특징(h_t^C)을 나타내는 (e)가 대상자의 얼굴 영상에서 감정을 인식할 수 있는 영역을 더 정밀하게 추출할 수 있음을 알 수 있다.

[0056] 이에 융합 컬러 디코더(330)의 인공 신경망 셀을 주의 가이드 LSTM(Attention Guided LSTM: 이하 AG-LSTM)이라 할 수 있다. 융합 컬러 디코더(330)또한 하나의 인공 신경망 셀(331)만을 구비하여 이전 출력을 다시 인가받는 시간에 따른 순환 구조로 구성될 수 있으나, 경우에 따라서는 다수(일예로 T개)의 인공 신경망 셀(331)을 포함하도록 구성될 수도 있다.

[0057] AG-LSTM으로 구현되는 융합 컬러 디코더(330)가 수행하는 기능은 수학식 2로 표현될 수 있다.

수학식 2

[0058]
$$
\begin{aligned}
i_t &= \sigma(w_{xi} * x_t^I + \sum_g w_{hi}^g * h_{t-1}^g + w_{ci} * c_{t-1} + b_i),\\
f_t &= \sigma(w_{xf} * x_t^I + \sum_g w_{hf}^g * h_{t-1}^g + w_{cf} * c_{t-1} + b_f),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(w_{xc} * x_t^I + \sum_g w_{hc}^g * h_{t-1}^g + b_c),\\
o_t &= \sigma(w_{xo} * x_t^I + \sum_g w_{ho}^g * h_{t-1}^g + w_{co} \odot c_t + b_o),\\
h_t^I &= o_t \odot \tanh(c_t)
\end{aligned}
$$

[0059] 여기서 $g \in \{I, D, F\}$이고, h_g 는 융합 컬러 디코더(330) 인공 신경망 셀인 AG-LSTM의 셀 출력으로서 컬러, 깊이 및 열을 포함하는 멀티모달 시퀀스 영상의 히든 특징을 나타낸다.

[0060] 도시하지 않았으나, 깊이 디코더(310)와 열 디코더(320) 및 융합 컬러 디코더(330)에는 순차적 디콘볼루션 레이어를 더 포함하여, 공간적 인코더(200)에서 축소된 공간 해상도를 다시 멀티모달 영상(I, D, F)의 크기로 확장시킬 수 있다.

[0061] 여기서는 설명의 편의를 위하여 인코더(200)와 시간적 디코더(300)를 구분하여 설명하였으나, 인코더(200)와 시간적 디코더(300)는 시공간 특징 추출부로 통합될 수 있다.

[0062] 또한 깊이 디코더(310)와 열 디코더(320)가 융합 컬러 디코더(330)의 컬러 특징맵(x^I) 시에 가이드 역할을 수행하는 깊이 시공간 특징(h^F)과 열 시공간 특징(h^F)을 제공하기 위한 구성이므로, 깊이 인코더(210)와 깊이 디코더(310)를 깊이 가이드 특징 추출부라 할 수 있으며, 열 인코더(220)와 열 디코더(320)를 열 가이드 특징 추출부라 할 수도 있다.

[0063] 시공간 주의 볼륨 획득부(400)는 멀티모달 영상 획득부(100)의 컬러 영상 획득부(130)로부터 컬러 시퀀스 영상($I_{1:T}$)을 인가받아 컬러 시퀀스 영상($I_{1:T}$)의 3D 특징 볼륨(X')을 추출하고, 시간적 디코더(300)의 융합 컬러 디코더(330)로부터 히든 특징을 인가받아 정규화하여 시공간 주의 볼륨(A)을 획득하며, 획득된 3D 특징 볼륨(X')과 시공간 주의 볼륨(A)을 결합하여, 주의 강화 특징 볼륨(X")을 획득한다.

[0064] 시공간 주의 볼륨 획득부(400)는 3D 특징 추출부(410), 정규화부(420) 및 주의 강화부(430)를 포함할 수 있다.

[0065] 3D 특징 추출부(410)는 컬러 영상 획득부(130)로부터 컬러 시퀀스 영상($I_{1:T}$)을 인가받아 특징을 추출한다. 3D 특징 추출부(410)는 공간적 인코더(200)의 컬러 인코더(230)와 유사하게 컬러 시퀀스 영상($I_{1:T}$)을 인가받아 특징을 추출하지만, 컬러 인코더(230)와 달리 컬러 시퀀스 영상($I_{1:T}$)의 각 프레임인 T개의 컬러 영상(I_1, I_2, …, I_T)을 개별적으로 인가받지 않고, T개의 컬러 영상(I_1, I_2, …, I_T)으로 구성되는 컬러 시퀀스 영상($I_{1:T}$) 전체를 인가받아 3차원의 단일 영상 개념으로 인식하여 3D 특징 볼륨(X')을 추출한다. 즉 시간의 흐름에 따라 누적된 T개의 2차원 컬러 영상(I_1, I_2, …, I_T)을 포함하는 컬러 시퀀스 영상($I_{1:T}$)을 3차원 이미지로 인식하여, 3차원의 컬러 시퀀스 영상(I1:T)에서 미리 학습된 방식에 따라 기법에 따라 특징을 추출함으로써 3D 특징 볼륨(X')을 획득한다. 따라서 3D 특징 볼륨(X')에도 T개의 컬러 영상(I_1, I_2, …, I_T) 각각의 공간적 특징과 T개의 컬러 영상(I_1, I_2, …, I_T) 사이의 관계에 대한 시간적 특징이 포함되는 것으로 볼 수 있다.

[0066] 여기서 3D 특징 추출부(410)는 일 예로 미리 학습된 3차원 콘볼루션 신경망(3D Convolutional Neural Networks: 이하 3D CNN)으로 구현될 수 있다.

[0067] 한편, 정규화부(420)는 시간적 디코더(300)의 융합 컬러 디코더(330)에서 순차적으로 출력되는 T개의 융합 컬러 시공간 특징(h_t^I)을 누적하여 시공간 특징 볼륨(H)으로 획득하고, 획득된 시공간 특징 볼륨(H)을 수학식 3에 따른 공간적 소프트 맥스(spatial softmax) 함수를 사용하여 정규화하여 시공간 주의 볼륨(A)을 획득한다.

수학식 3

$$A_{t,i} = \frac{\exp(H_{t,i})}{\sum_j \exp(H_{t,j})}$$

[0068]

[0069] 여기서 $H_{t,i}$ 는 시공간 특징 볼륨(H)에 포함된 시간(t)에서의 융합 컬러 시공간 특징(h_t^I)인 시공간 특징맵(H_t)에서 i(i ∈ {1, ㅇㅇㅇ , H ㅅ W}) 픽셀 위치의 특징을 나타내고, $A_{t,i}$ 는 시공간 특징 볼륨(H)의 위치별 특징($H_{t,i}$)에 대응하는 시공간 주의 볼륨(A)의 가중치를 나타낸다.

[0070] 정규화부(420)는 일 예로 시공간 주의 볼륨(A)에 포함되는 모든 가중치($A_{t,i}$)의 합이 1이되도록 정규화를 수행할 수 있다.

[0071] 주의 강화부(430)는 3D 특징 추출부(410)에서 획득된 3D 특징 볼륨(X')과 정규화부(420)에서 획득된 시공간 주의 볼륨(A)을 인가받아 수학식 3과 같이 3D 특징 볼륨(X')과 3D 특징 볼륨(X')을 하다마드 곱셈하여 주의 강화 특징 볼륨(X")을 획득한다.

수학식 4

$$X'' = A \odot X'$$

[0072]

[0073] 여기서 ⊙는 하다마드(Hadamard) 곱셈 연산자를 나타낸다.

[0074] 감정 추정부(500)는 미리 학습된 인공 신경망으로 구현되어 미리 학습된 방식에 따라 시공간 주의 볼륨 획득부(400)로부터 인가되는 주의 강화 특징 볼륨(X")으로부터 감정을 추정하여 감정값(y)을 획득한다. 여기서 감정 추정부(500)는 3D CNN으로 구현될 수 있으며, 감정 추정부(500)는 도 3의 우측단에 도시된 바와 같이, 발란스와 흥분도로 구성되는 2차원 감정 공간 상에 특정 포인트에 대한 좌표 값으로, -1 에서 1 사이의 스칼라 값(scalar value)([-1, 1])으로 획득되도록 감정값(y)을 추출할 수 있다.

[0075] 한편, 본 실시예에 따른 감정 인식 장치(10)는 학습을 수행하기 위한 학습부(미도시)를 더 포함할 수 있다.

[0076] 학습부는 감정값(y)과 다수의 대상자를 대상으로 설문 등으로 방식으로 실제 측정한 진리값($\hat{y}$)을 비교하여 수학식 5에 따라 손실(L)을 계산하고, 계산된 손실(L)을 역전파하여 감정 인식 장치(10)의 인공 신경망을 학습시킬 수 있다.

수학식 5

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M} \|\hat{y}_m - y_m\|_2$$

[0077]

[0078] 여기서 $\|\cdot\|_2$ 는 L2-norm 함수이다.

[0079] 결과적으로 본 실시예에 따른 감정 인식 장치는 컬러 영상과 깊이 영상 및 열 영상을 포함하는 멀티모달 영상을 인가받고, 멀티모달 영상을 융합하여 감정을 인식함으로써 매우 정확하게 감정을 인식할 수 있으며, 영상에 포함된 다수 프레임에 대한 시간적 변화가 함께 반영되도록 하여 감정 인식 정확도를 더욱 향상시킬 수 있다.

[0080] 도 6은 본 발명의 일 실시예에 따른 감정 인식 방법을 나타낸다.

[0081] 도 2 내지 도 5를 참조하여, 도 6의 감정 인식 방법을 설명하면, 우선 깊이 시퀀스 영상($D_{1:T}$)과 열 시퀀스 영상($F_{1:T}$) 및 컬러 시퀀스 영상($I_{1:T} = \{I_1, I_2, \cdots, I_T\}$)이 포함되는 멀티모달 영상을 획득한다(S10).

[0082] 그리고 미리 학습된 인공 신경망으로 구현되는 인코더를 이용하여 멀티모달 영상에 포함된 깊이 시퀀스 영상($D_{1:T}$)과 열 시퀀스 영상($F_{1:T}$) 및 컬러 시퀀스 영상($I_{1:T} = \{I_1, I_2, \cdots, I_T\}$) 각각에서 깊이 영상과 열영상 및 컬러 영상을 학습되는 방식에 따라 공간적 특징을 추출하도록 순차적으로 인코딩하여 각각 K개의 깊이 특징맵(x^D)과 K개의 열 특징맵(x^F) 및 K개의 컬러 특징맵(x^I)을 시간 순서에 따라 획득한다(S20).

[0083] 이후, 미리 학습된 인공 신경망으로 구현되는 디코더를 이용하여 K개의 깊이 특징맵(x^D)과 K개의 열 특징맵(x^F) 각각을 시간 순서에 따라 순차적으로 인가받고, 순차적으로 인가되는 K개의 깊이 특징맵(x^D)과 K개의 열 특징맵(x^F) 각각에 대해 시간적 특징이 반영되도록 디코딩하여, 깊이 시공간 특징(h^F)과 열 시공간 특징(h^F)을 획득한다(S30).

[0084] 그리고 K개의 컬러 특징맵(x^I)과 함께 대응하는 깊이 시공간 특징(h^F)과 열 시공간 특징(h^F)을 시간 순서에 따라 순차적으로 인가받아 시간적 특징이 반영되도록 디코딩하여, 융합 컬러 시공간 특징(h^I)을 획득한다(S40).

[0085] 그리고 컬러 시퀀스 영상($I_{1:T} = \{I_1, I_2, \cdots, I_T\}$) 전체에 대해 미리 학습된 인공 신경망을 이용하여 3D 특징(X')을 추출한다(S50).

[0086] 그리고 융합 컬러 시공간 특징(h^I)을 누적하고, 기지정된 방식으로 정규화하여 시공간 주의 볼륨(A)을 획득하고, 획득된 시공간 주의 볼륨(A)과 3D 특징(X')을 기지정된 방식으로 결합하여, 주의 강화 특징 볼륨(X")을 획득한다(S60).

[0087] 주의 강화 특징 볼륨(X")이 획득되면, 미리 학습된 인공 신경망을 이용하여 주의 강화 특징 볼륨(X")으로부터 감정을 추정하여 감정값(y)을 획득한다(S70).

[0088] 한편, 현재 단계가 학습 단계인지 판별한다(S80). 만일 현재 단계가 학습 단계인 것으로 판별되면, 미리 획득된 진리값($\hat{y}$)과 감정값(y)을 비교하여 수학식 5에 따라 손실(L)을 계산하고, 계산된 손실을 역전파하여 학습을 수행한다(S90). 여기서 학습은 기지정된 횟수 또는 손실(L)이 기지정된 문턱값 이하가 될때까지 반복하여 수행될 수 있다.

[0089] 본 발명에 따른 방법은 컴퓨터에서 실행시키기 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.

[0090] 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.

[0091] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0092] 10: 감정 인식 장치 100: 멀티모달 영상 획득부

110: 깊이 영상 획득부 120: 열 영상 획득부

130: 컬러 영상 획득부 200: 공간적 인코더

210: 깊이 인코더 220: 열 인코더

230: 컬러 인코더 300: 시간적 디코더

310: 깊이 디코더 320: 열 디코더

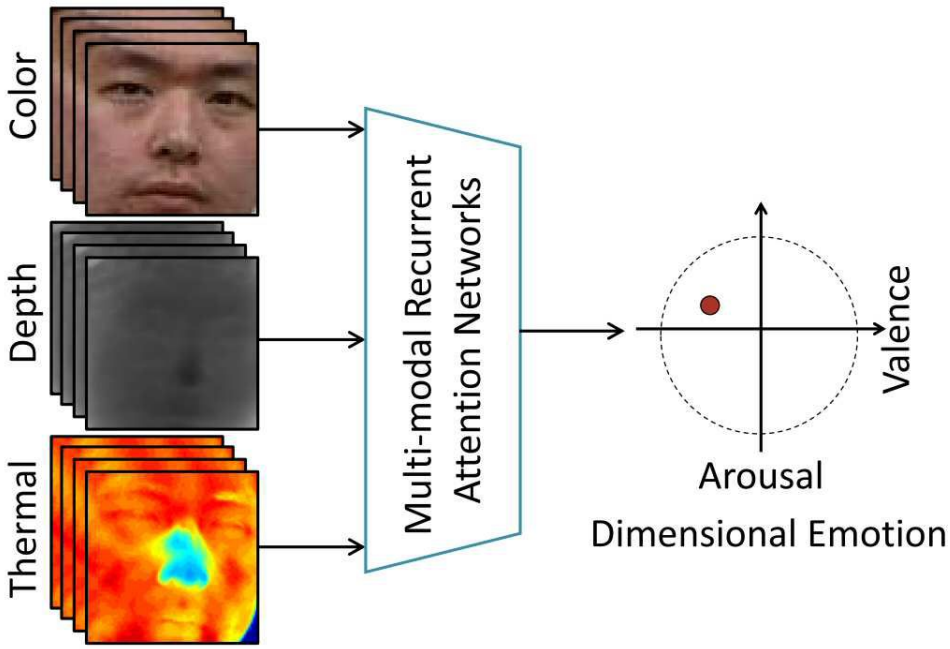
330: 융합 컬러 인코더 400: 시공간 주의 볼륨 획득부

410: 3D 특징 추출부 420: 정규화부

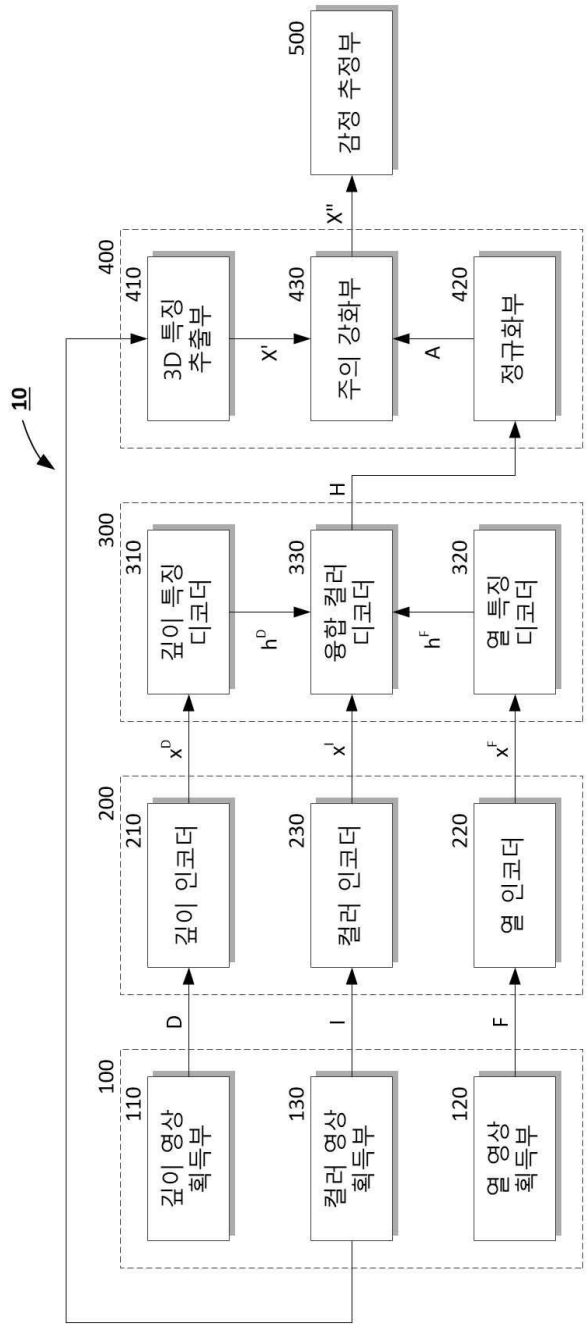
430: 주의 강화부 500: 감정 추정부

도면

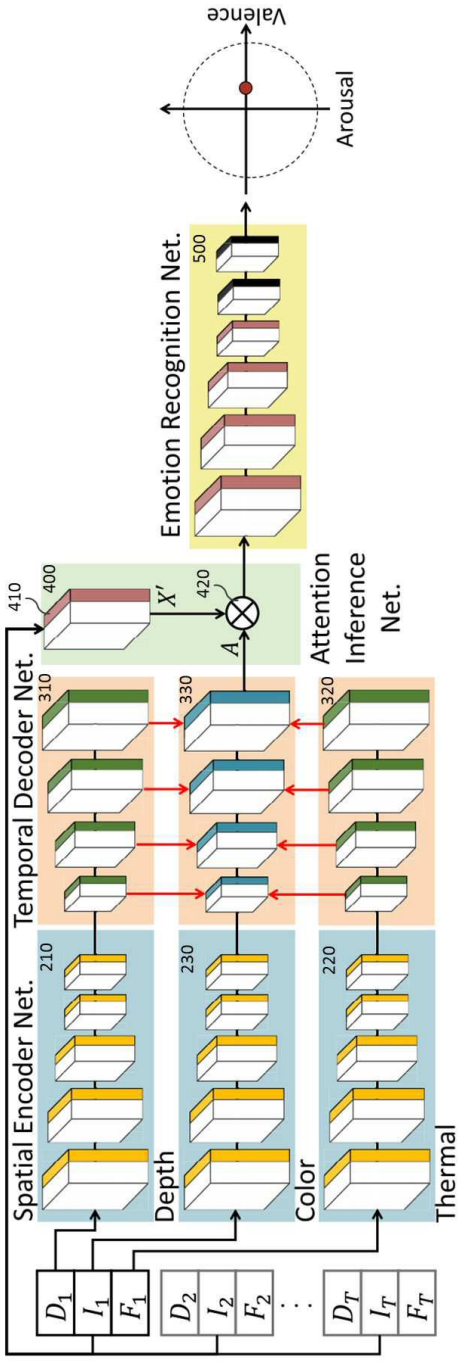
도면1



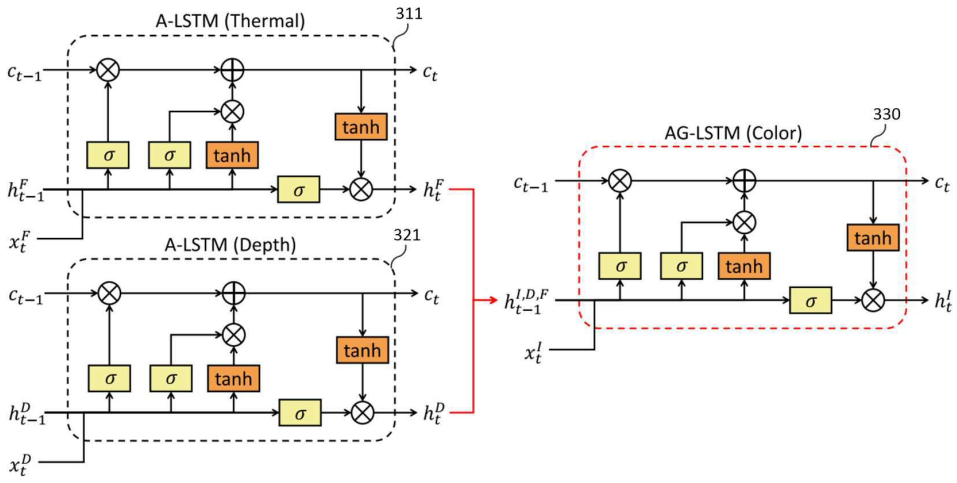
도면2


10
100
110
깊이 영상 획득부
130
컬러 영상 획득부
120
열 영상 획득부
D
I
F
200
210
깊이 인코더
230
컬러 인코더
220
열 인코더
x^D
x^I
x^F
300
310
깊이 특징 디코더
330
융합 컬러 디코더
320
열 특징 디코더
h^D
h^F
H
400
410
3D 특징 추출부
430
주의 강화부
420
정규화부
X'
A
X''
500
감정 추정부

도면3

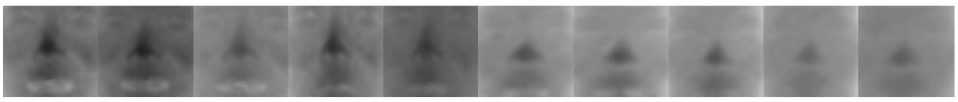

Valence
Arousal
Emotion Recognition Net.
500
Attention Inference Net.
410
400
X'
420
A
Temporal Decoder Net.
310
330
320
Spatial Encoder Net.
210
230
220
Depth
Color
Thermal
D1
I1
F1
D2
I2
F2
DT
IT
FT

도면4

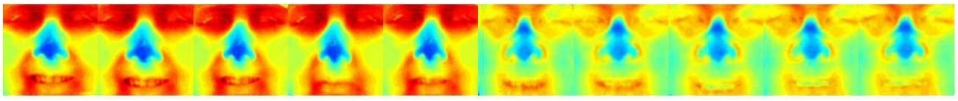


도면5

(a) Color sequence



(b) Depth sequence

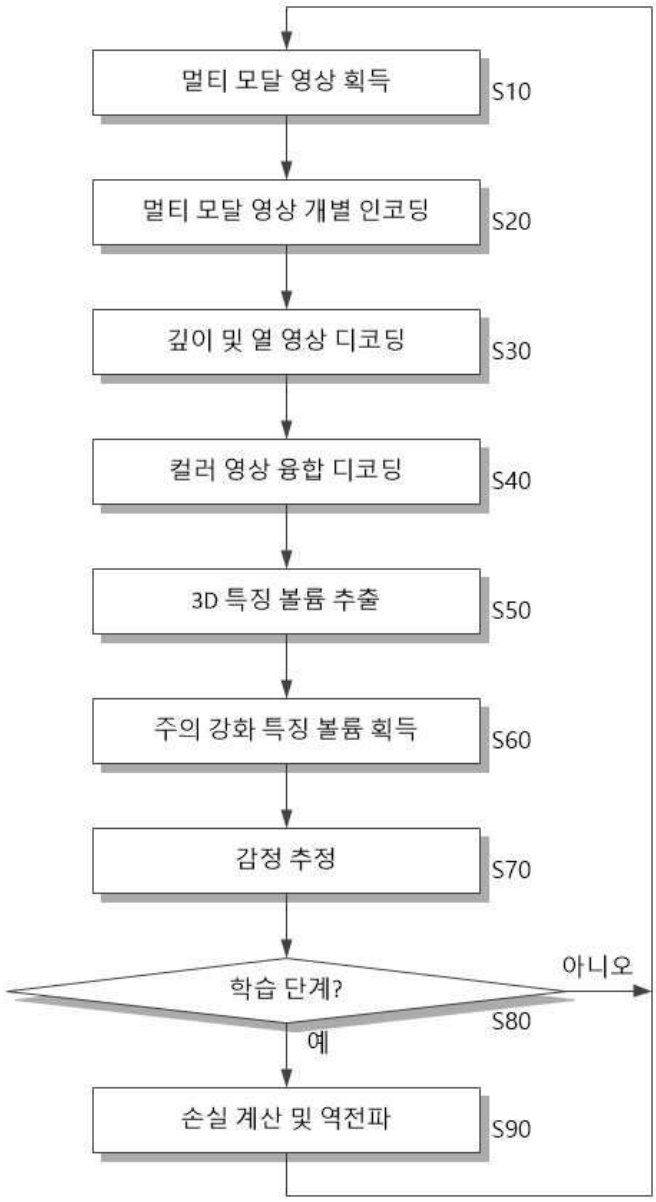


(c) Thermal sequence

(d) Attention scores without using depth and thermal sequences

(e) Attention scores

도면6


멀티 모달 영상 획득
S10
멀티 모달 영상 개별 인코딩
S20
깊이 및 열 영상 디코딩
S30
컬러 영상 융합 디코딩
S40
3D 특징 볼륨 추출
S50
주의 강화 특징 볼륨 획득
S60
감정 추정
S70
학습 단계?
S80
아니오
예
손실 계산 및 역전파
S90