(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년07월09일
(11) 등록번호 10-2275436
(24) 등록일자 2021년07월05일

(51) 국제특허분류(Int. Cl.)
A61B 5/16 (2006.01) *A61B 5/00* (2021.01)
A61B 5/24 (2021.01) *A61B 5/369* (2021.01)
G06K 9/00 (2006.01) *G16H 50/20* (2018.01)
(52) CPC특허분류
A61B 5/165 (2013.01)
A61B 5/316 (2021.01)
(21) 출원번호 10-2020-0007223
(22) 출원일자 2020년01월20일
심사청구일자 2020년01월20일
(56) 선행기술조사문헌
Choi D Y, etc., Multimodal Attention Network for Continuous-Time Emotion Recognition Using Video and EEG Signals. IEEE Access. Vol.8, pp.203814~203826 (2020.11.09.)
Soleymani M, etc., Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Transactions on Affective Computing. Vol.7, No.1, pp.17~28 (March, 2016)

(73) 특허권자
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
송병철
서울특별시 양천구 목동서로 38, 125동 307호(목동, 목동신시가지아파트1단지)
최동윤
충청남도 서산시 고북면 고북1로 299, 204호
(74) 대리인
양성보

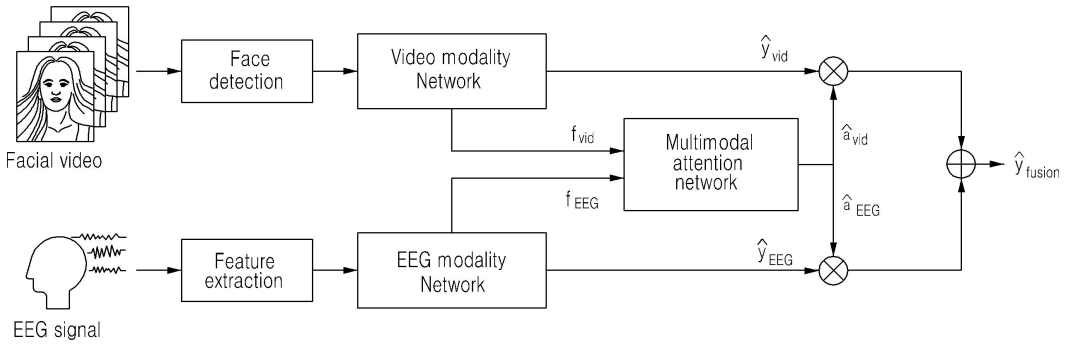
전체 청구항 수 : 총 11 항 심사관 : 유창용

(54) 발명의 명칭 **실시간 감정인식을 위한 영상 및 EEG 신호의 융합 기술**

(57) 요 약

실시간 감정인식을 위한 영상 및 EEG 신호의 융합 기술이 개시된다. 일 실시예에 따른 감정 인식 시스템에 의해 수행되는 감정 인식 방법은, 비디오 신호와 EEG 신호로부터 각각의 모달리티 특징을 추출하는 단계; 상기 추출된 각각의 모달리티 특징을 멀티모달 어텐션 네트워크(Multimodal attention network)에 입력하여 모달리티에 대한 어텐션 가중치를 결정하는 단계; 및 상기 결정된 어텐션 가중치를 상기 추출된 각각의 모달리티 특징에 반영하여 융합된 감정 정보를 출력하는 단계를 포함할 수 있다.

대 표 도



(52) CPC특허분류
A61B 5/369 (2021.01)
A61B 5/7275 (2013.01)
A61B 5/7278 (2021.01)
G06K 9/00302 (2013.01)
G16H 50/20 (2018.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1415162019
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 산업핵심기술개발사업
연구과제명 인간 내면상태의 인식 및 이를 이용한 인간친화형 인간-로봇 상호작용 기술 개발(3차년도)
기 여 율 1/1
과제수행기관명 인하대학교 산학협력단
연구기간 2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

감정 인식 시스템에 의해 수행되는 감정 인식 방법에 있어서,

비디오 신호와 EEG 신호로부터 각각의 모달리티 특징을 추출하는 단계;

상기 추출된 각각의 모달리티 특징을 멀티모달 어텐션 네트워크(Multimodal attention network)에 입력하여 모달리티에 대한 어텐션 가중치를 결정하는 단계; 및

상기 결정된 어텐션 가중치를 상기 추출된 각각의 모달리티 특징에 반영하여 융합된 감정 정보를 출력하는 단계

를 포함하는 감정 인식 방법.

청구항 2

제1항에 있어서,

상기 결정하는 단계는,

상기 비디오 신호로부터 추출된 제1 비디오 모달리티 특징과 상기 EEG 신호로부터 추출된 제1 EEG 모달리티 특징을 융합하여 제1 융합 특징을 생성하고, 상기 제1 비디오 모달리티 특징 및 상기 제1 EEG 모달리티 특징을 상기 멀티모달 어텐션 네트워크의 완전 연결 레이어를 통과시킴에 따라 제2 비디오 모달리티 특징 및 제2 EEG 모달리티 특징을 획득하는 단계

를 포함하는 감정 인식 방법.

청구항 3

제2항에 있어서,

상기 결정하는 단계는,

상기 생성된 제1 융합 특징과 상기 제2 비디오 모달리티 특징 및 상기 제2 EEG 모달리티 특징을 융합하여 제2 융합 특징을 획득하고, 상기 획득된 제2 융합 특징을 이용하여 각각의 모달리티의 어텐션을 계산하는 단계

를 포함하는 감정 인식 방법.

청구항 4

제1항에 있어서,

상기 멀티모달 어텐션 네트워크는, 하위 계층 분해(low-rank decomposition) 기반 쌍선형 풀링(bilinear pooling)을 멀티-레이어(multi-layer)로 구성되고, 상기 각각의 모달리티 특징을 융합시킴에 따라 획득된 발란스 정보는, 상기 멀티모달 어텐션 네트워크를 통해 각각 출력된 정보가 어텐션 가중치로 가중합(weight sum)되어 출력되는, 감정 인식 방법.

청구항 5

제1항에 있어서,

상기 추출하는 단계는,

입력 영상 시퀀스에서 크롭(crop)된 얼굴 이미지를 획득하고, 상기 획득된 얼굴 이미지를 딥 컨볼루션 인코더(deep convolution encoder)를 이용하여 비디오 모달리티 특징을 추출하고, 상기 추출된 비디오 모달리티 특징을 비디오 모달리티 네트워크를 통해 회귀(regression)를 수행하는 단계

를 포함하는 감정 인식 방법.

청구항 6

제5항에 있어서,

상기 추출하는 단계는,

상기 획득된 얼굴 이미지를 CNN 기반의 딥 컨볼루션 인코더를 이용하여 비디오 신호를 1차원의 특징 벡터로 변환하고, 상기 변환된 1차원의 특징 벡터를 비디오 모달리티 네트워크를 통과시킴에 따라 출력되는 히든 스테이트 벡터(hidden state vector)를 완전 연결 레이어에 통과시켜 입력 영상 시퀀스의 발란스 라벨 값을 출력하는 단계

를 포함하고,

상기 비디오 모달리티 네트워크는, LSTM 네트워크를 포함하고,

상기 변환된 1차원의 특징 벡터를 상기 LSTM 네트워크를 통과시키기 전에 완전 연결 레이어를 통과시켜 특징 벡터의 차원이 조정되는, 감정 인식 방법.

청구항 7

제1항에 있어서,

상기 추출하는 단계는,

상기 EEG 신호로부터 시간 영역, 주파수 영역 및 시간-주파수 영역에 대한 EEG 모달리티 특징을 추출하고, 상기 추출된 EEG 모달리티 특징을 EEG 모달리티 네트워크에 입력하여 EEG 신호의 발란스 라벨 값을 출력하는 단계

를 포함하고,

상기 EEG 모달리티 네트워크는, LSTM 네트워크를 포함하는, 감정 인식 방법.

청구항 8

제2항에 있어서,

상기 결정하는 단계는,

상기 추출된 제1 비디오 모달리티 특징과 상기 추출된 제1 EEG 모달리티 특징을 하위 계층 분해를 적용함에 따라 획득된 가중치들을 각각의 모달리티 특징 벡터에 적용시켜 멀티곱을 통해 융합시키는 단계

를 포함하는 감정 인식 방법.

청구항 9

제3항에 있어서,

상기 결정하는 단계는,

상기 제1 융합 특징과 상기 제2 비디오 모달리티 특징 및 상기 제2 EEG 모달리티 특징을 하위 계층 분해(low-rank decomposition)에 기반하여 융합시키는 단계

를 포함하는 감정 인식 방법.

청구항 10

제1항에 있어서,

상기 결정하는 단계는,

상기 비디오 신호를 비디오 모달리티 네트워크를 통해 추출된 제1 비디오 모달리티 특징, 상기 EEG 신호를 EEG 모달리티 네트워크를 통해 추출된 제1 EEG 모달리티 특징을 멀티모달 어텐션 네트워크에 입력함에 따라 출력된 어텐션 가중치를 어텐션 라벨과 오차를 측정하여 트레이닝 손실을 계산하는 단계

를 포함하는 감정 인식 방법.

청구항 11

감정 인식 시스템에 있어서,

비디오 신호와 EEG 신호로부터 각각의 모달리티 특징을 추출하는 추출부;

상기 추출된 각각의 모달리티 특징을 멀티모달 어텐션 네트워크(Multimodal attention network)에 입력하여 모달리티에 대한 어텐션 가중치를 결정하는 결정부; 및

상기 결정된 어텐션 가중치를 상기 추출된 각각의 모달리티 특징에 반영하여 융합된 감정 정보를 출력하는 출력부

를 포함하는 감정 인식 시스템.

발명의 설명

기 술 분 야

[0001] 아래의 설명은 영상 및 EEG 신호를 융합한 실시간 감정 인식 기술에 관한 것이다.

배 경 기 술

[0003] 인간의 감정을 인식하는 기술은 궁극적인 로봇과 인간의 상호작용을 위한 핵심 기술이다. 또한 감정인식은 인공지능 분야에서 최근 많은 관심을 받고 있다. 현재까지 개발된 감정인식 기술들을 살펴보면, 주로 얼굴 이미지(facial image)로부터 얻어진 특징들에 기반하여 표정의 변화를 인지함으로써 감정의 카테고리를 구분하는 방식들이 지배적이다. 최근에는 별도의 특징 추출 과정없이 convolution neural network(CNN)을 이용하여 엔드-투-엔드(end-to-end)로 감정을 분류(classification)하는 메커니즘(mechanism)들이 개발되어 높은 성능을 보이고 있다.

[0004] 또한, 음성 신호로부터 추출된 일종의 톤(tone) 정보로부터 인간의 감정을 인지하려는 시도들도 있었다. 그러나 음성 정보는 드문(sparse)하게 존재하기 때문에 연속적인 감정을 추출하는데 근본적인 한계가 있다. 최근에는 인간의 뇌에서 발생하는 전기적 생체 신호인 electroencephalogram(EEG)을 이용하여 감정 인식을 수행하는 연구들도 진행되고 있다. 예를 들면, 파워 스펙트럼 밀도(power spectral density)와 같은 주파수 영역 특징(frequency domain feature)을 추출한 후 전형적인 기계학습 알고리즘을 적용하여 감정을 인식할 수 있다. 최근 EEG 신호의 채널 간 비대칭(asymmetry) 특성을 특징(feature)으로 추출하고, 여기에 advanced deep learning 알고리즘을 적용하여 감정인식의 정확도를 향상시킨 사례가 있다.

[0005] 한편, 영상 신호, 음성 신호, 생체 신호 중 두 개 이상을 동시에 이용하는 소위 멀티-모달(multi-modal) 신호에 기반한 감정인식 기술들이 단일 신호 기반 방식들보다 우수하다는 연구 결과들이 보고되고 있다. 비특허문헌 1 "Soleymani et. al."에 따르면, EEG 신호와 얼굴 랜드마크(facial landmark)로 구성된 멀티-모달 데이터에 long short-term memory(LSTM)을 적용한 continuous-time valence domain의 emotion regression 방법을 제안하였다. 그러나 상기 기법들은 입력data concatenation이나 출력 평균(average) 등 비교적 단순한 방식의 모달리티 퓨전(modality fusion)이기 때문에 모달리티 간 상호보완을 효과적으로 이용하지 못하는 단점이 있었다.

[0006] 비특허문헌 1: M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. "Analysis of EEG signals and facial expressions for continuous emotion detection," IEEE Transactions on Affective Computing, vol. 7 no. 1, pp. 17-28, Mar. 2016.

발명의 내용

해결하려는 과제

[0007] 비디오 데이터와 EEG 신호를 포함하는 멀티 모달 신호를 상호 보완하여 시너지를 내는 모달리티 어텐션 기반의 융합 방법 및 시스템을 제공할 수 있다.

[0008] 비디오 데이터와 EEG 신호에 대한 각각의 특징을 융합하고 모달리티 어텐션을 결정하는 멀티모달 어텐션 네트워크를 제공할 수 있다.

과제의 해결 수단

[0010] 감정 인식 시스템에 의해 수행되는 감정 인식 방법은, 비디오 신호와 EEG 신호로부터 각각의 모달리티 특징을 추출하는 단계; 상기 추출된 각각의 모달리티 특징을 멀티모달 어텐션 네트워크(Multimodal attention network)에 입력하여 모달리티에 대한 어텐션 가중치를 결정하는 단계; 및 상기 결정된 어텐션 가중치를 상기 추출된 각각의 모달리티 특징에 반영하여 융합된 감정 정보를 출력하는 단계를 포함할 수 있다.

[0011] 상기 결정하는 단계는, 상기 비디오 신호로부터 추출된 제1 비디오 모달리티 특징과 상기 EEG 신호로부터 추출된 제1 EEG 모달리티 특징을 융합하여 제1 융합 특징을 생성하고, 상기 제1 비디오 모달리티 특징 및 상기 제1 EEG 모달리티 특징을 상기 멀티모달 어텐션 네트워크의 완전 연결 레이어를 통과시킴에 따라 제2 비디오 모달리티 특징 및 제2 EEG 모달리티 특징을 획득하는 단계를 포함할 수 있다.

[0012] 상기 결정하는 단계는, 상기 생성된 제1 융합 특징과 상기 제2 비디오 모달리티 특징 및 상기 제2 EEG 모달리티 특징을 융합하여 제2 융합 특징을 획득하고, 상기 획득된 제2 융합 특징을 이용하여 각각의 모달리티의 어텐션을 계산하는 단계를 포함할 수 있다.

[0013] 상기 멀티모달 어텐션 네트워크는, 하위 계층 분해(low-rank decomposition) 기반 쌍선형 풀링(bilinear pooling)을 멀티-레이어(multi-layer)로 구성되고, 상기 각각의 모달리티 특징을 융합시킴에 따라 획득된 발란스 정보는, 상기 멀티모달 어텐션 네트워크를 통해 각각 출력된 정보가 어텐션 가중치로 가중합(weight sum)되어 출력될 수 있다.

[0014] 상기 추출하는 단계는, 입력 영상 시퀀스에서 크롭(crop)된 얼굴 이미지를 획득하고, 상기 획득된 얼굴 이미지를 딥 컨볼루션 인코더(deep convolution encoder)를 이용하여 비디오 모달리티 특징을 추출하고, 상기 추출된 비디오 모달리티 특징을 비디오 모달리티 네트워크를 통해 회귀(regression)를 수행하는 단계를 포함할 수 있다.

[0015] 상기 추출하는 단계는, 상기 획득된 얼굴 이미지를 CNN 기반의 딥 컨볼루션 인코더를 이용하여 비디오 신호를 1차원의 특징 벡터로 변환하고, 상기 변환된 1차원의 특징 벡터를 비디오 모달리티 네트워크를 통과시킴에 따라 출력되는 히든 스테이트 벡터(hidden state vector)를 완전 연결 레이어에 통과시켜 입력 영상 시퀀스의 발란스 라벨 값을 출력하는 단계를 포함하고, 상기 비디오 모달리티 네트워크는, LSTM 네트워크를 포함하고, 상기 변환된 1차원의 특징 벡터를 상기 LSTM 네트워크를 통과시키기 전에 완전 연결 레이어를 통과시켜 특징 벡터의 차원이 조정될 수 있다.

[0016] 상기 추출하는 단계는, 상기 EEG 신호로부터 시간 영역, 주파수 영역 및 시간-주파수 영역에 대한 EEG 모달리티 특징을 추출하고, 상기 추출된 EEG 모달리티 특징을 EEG 모달리티 네트워크에 입력하여 EEG 신호의 발란스 라벨 값을 출력하는 단계를 포함하고, 상기 EEG 모달리티 네트워크는, LSTM 네트워크를 포함할 수 있다.

[0017] 상기 결정하는 단계는, 상기 추출된 제1 비디오 모달리티 특징과 상기 추출된 제1 EEG 모달리티 특징을 하위 계층 분해를 적용함에 따라 획득된 가중치들을 각각의 모달리티 특징 벡터에 적용시켜 멀티곱을 통해 융합시키는 단계를 포함할 수 있다.

[0018] 상기 결정하는 단계는, 상기 제1 융합 특징과 상기 제2 비디오 모달리티 특징 및 상기 제2 EEG 모달리티 특징을 하위 계층 분해(low-rank decomposition)에 기반하여 융합시키는 단계를 포함할 수 있다.

[0019] 상기 결정하는 단계는, 상기 비디오 신호를 비디오 모달리티 네트워크를 통해 추출된 제1 비디오 모달리티 특징, 상기 EEG 신호를 EEG 모달리티 네트워크를 통해 추출된 제1 EEG 모달리티 특징을 멀티모달 어텐션 네트워크에 입력함에 따라 출력된 어텐션 가중치를 어텐션 라벨과 오차를 측정하여 트레이닝 손실을 계산하는 단계를 포함할 수 있다.

발명의 효과

[0021] 영상기반 네트워크와 EEG 기반 네트워크를 융합하는 기술로 모달리티 어텐션 네트워크를 제공함에 따라 영상 정보 혹은 EEG 정보 각각의 단일 모달리티 대비 개선된 감정 인식 성능을 제공할 수 있다

[0022] 모달리티 어텐션 네트워크에서 각 네트워크로부터 출력되는 중간 정보를 이용하여 감정상태 추정 정확도가 높은 모달리티를 예측하고 해당 모달리티의 결과를 선택하여 출력할 수 있다. 이에 따라 각 모달리티가 감정 인식에 유리한 신호에 대해서만 처리된 결과를 선택함으로써 성능을 향상시킬 수 있다.

[0023] 또한, 모달리티 어텐션 네트워크에서 하위 계층 분해(low rank decomposition) 기반의 쌍선형 풀링(bilinear pooling)을 적용하고 이를 멀티 레이어(multi layer)로 적용함으로써 융합 과정을 효과적으로 수행할 수 있다.

[0024] 또한, 기 설정된 시간(예를 들면, 약 2초 분량)의 짧은 시퀀스(sequence) 정보를 이용함으로써 연속적인 시간에서의 감정 인식이 가능하다.

도면의 간단한 설명

[0026] 도 1은 일 실시예에 따른 감정 인식 시스템에서 영상 신호와 EEG 신호를 융합하여 감정을 인식하는 동작을 설명하기 위한 도면이다.

도 2는 일 실시예에 따른 감정 인식 시스템에서 모달리티 어텐션 네트워크를 설명하기 위한 도면이다.

도 3은 일 실시예에 따른 감정 인식 시스템에서 긍정/부정을 추정한 결과를 나타낸 예이다.

도 4는 일 실시예에 따른 감정 인식 시스템의 구성을 설명하기 위한 블록도이다.

도 5는 일 실시예에 따른 감정 인식 시스템에서 감정 인식 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0027] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0029] 도 1은 일 실시예에 따른 감정 인식 시스템에서 영상 신호와 EEG 신호를 융합하여 감정을 인식하는 동작을 설명하기 위한 도면이다.

[0030] 감정 인식 시스템은 비디오 모달리티(Video modality) 및 EEG 모달리티(EEG modality)의 네트워크로 처리된 중간의 특징 정보를 융합(fusion)하고 입력에 대한 모달리티 어텐션 가중치(modality attention weight)를 측정할 수 있다. 감정 인식 시스템은 각 네트워크로부터 출력되는 결과에 모달리티 어텐션 가중치를 반영하여 융합된 감정 정보를 출력할 수 있다. 이때, 입력 신호의 특성을 분석하여 비디오 모달리티와 EEG 모달리티에 대한 유효성을 확인하고 두 출력의 가중치를 조절할 수 있다면 비디오와 EEG를 동시에 이용함으로써 신호 모달리티 대비 감정 인식의 성능을 개선할 수 있다.

[0031] 감정 인식 시스템은 각 모달리티의 특징을 융합하기 위하여 하위 계층 분해(low-rank decomposition) 기반 쌍선형 풀링(bilinear pooling)을 멀티 레이어(multi-layer)로 구성한 멀티모달 어텐션 네트워크를 제안할 수 있다. 적은 연산량 및 메모리를 이용하면서 멀티모달 특징을 효과적으로 융합하고 어텐션 추정 및 감정 인식 성능을 개선시킬 수 있다.

[0032] 감정 인식 시스템은 비디오 모달리티 네트워크를 구성할 수 있다. 이때, 비디오 모달리티 네트워크는 영상(이미지) 신호 기반의 네트워크일 수 있다. 감정 인식 시스템은 얼굴 검출 알고리즘을 이용하여 입력 영상 시퀀스에서 크롭(Crop)된 얼굴 이미지들을 획득할 수 있다. 딥 컨볼루션 인코더(deep convolution encoder)를 통해 특징을 추출한 후, LSTM을 통해 회귀를 수행하는 비디오 모달리티 네트워크가 동작될 수 있다. 비디오 모달리티 네트워크는 CNN-LSTM 기반 네트워크의 일반적인 구조로 구성될 수 있으며, 일반적으로 비디오 입력에 대한 인식을 수행할 수 있다.

[0033] 감정 인식 시스템은 2차원의 얼굴 이미지 정보를 LSTM의 입력으로 이용하기 위하여, CNN 기반의 딥 컨볼루션 인코더를 이용하여 2차원의 얼굴 이미지 정보를 1차원의 벡터 형태로 변환함에 따라 1차원의 특징 벡터를 추출할 수 있다. 이때, CNN 기법 중 매우 우수한 성능을 보이는 DenseNet을 기반으로 딥 컨볼루션 인코더를 설계할 수 있다. 입력 영상의 크기는 224x224 이며, DesneNet의 'fc2 레이어'를 통과한 4096길이의 특징 벡터가 출력될 수 있다.

[0034] 감정 인식 시스템은 시퀀스 내 각 영상에서 추출된 1차원의 특징 벡터를 이용하여 하나의 발란스(valence) 값을 회귀(regression)하기 위하여 널리 알려진 시간의 네트워크(temporal network)인 LSTM 네트워크를 이용할 수 있다. 딥 컨볼루셔널 특징들이 LSTM으로 처리되기 전, 완전 연결 레이어(fully connected layer)를 통과하여 특징 벡터의 차원으로 조정하고, LSTM에서 출력되는 히든 스테이트 벡터(hidden state vector)가 한 번 더 완전 연결 레이어에 통과되어 시퀀스의 발란스 라벨(valence label) 값을 출력할 수 있다.

[0035] 감정 인식 시스템은 EEG 신호 기반의 네트워크를 구성할 수 있다. 먼저, 윈도윙(windowing)을 수행한 EEG 신호에 대하여 시간, 주파수, 시간-주파수 영역에 대한 특징을 추출하여 1차원의 EEG 특징을 추출할 수 있다. 시간

에 따른EEG 특징의 시퀀스를 이용하여 LSTM으로 구성된 네트워크에 기반한 회귀를 수행할 수 있다.

[0036] 구체적으로, 감정 인식 시스템은 EEG 신호로부터 EEG 특징을 추출할 수 있다. 기존 연구에서는 EEG의 공간분해능의 우수함 때문에 주파수 영역의 특징이 많이 사용되었다. 실시예에서는 주파수 영역의 특징뿐만 아니라, 시간, 시간-주파수 영역의 특징을 포함시킬 수 있다. 평균(mean), 최소(minimum), 최대(maximum), 1차 차이(1st difference), 정규화 된 1차 차이(normalized 1st difference)의 특징은 시간에 따라 감정이 바뀌었을 때는 인식할 수 있다. 주파수 영역에서는 SD(Power spectrum density)를 느린 알파(Slow Alpha, 8-10Hz), 알파(Alpha, 8-12.9Hz), 베타(Beta, 13-29.9Hz), 감마(Gamma, 30-50Hz) 4가지 대역 영역으로 구분할 수 있다. 시간-주파수 영역은 주파수 영역과 같이 4가지의 대역 영역으로 구분될 수 있으며, 5가지의 특징을 선택할 수 있다. 이산 웨이브렛 변환(DWT, Discrete wavelet transform)는 신호를 시간을 기준으로 하여 다양한 비트로 분해할 수 있다. 이산 웨이브렛 변환의 각 주파수 영역에서mean, max, abs value를 사용했으며, 이산 웨이브렛 변환에서 Log, Abs(Log) value를 추가로 사용할 수 있다.

[0037] 감정 인식 시스템은 EEG 신호에 대하여 EEG 특징을 추출함에 따라 1차원의 특징으로 구성된 시퀀스를 생성할 수 있다. 영상 모달리티와 동일하게 LSTM 기반의 회귀를 이용하여 발란스 라벨(valence label)값을 출력할 수 있다.

[0038] 도 2를 참고하면, 멀티모달 어텐션 네트워크를 설명하기 위한 도면이다. 감정 인식 시스템은 비디오 모달리티 및 EEG 모달리티의 각 네트워크에서 추출된 중간 특징을 이용하여 모달리티에 대한 어텐션 가중치를 출력할 수 있다. 이때, 특징 융합 과정을 반복적으로 수행하여 융합된 특징(fused feature)을 개선(refine)하는 네트워크를 제공할 수 있다.

[0039] 감정 인식 시스템은 비디오 특징과 EEG 특징(제1 모달리티 특징) f_{vid}, f_{EEG} 를 융합하여 융합된 특징(제1 융합 특징) $f_{fusion1}$을 생성할 수 있다. 감정 인식 시스템은 융합된 특징(제1 융합 특징) $f_{fusion1}$과 각 모달리티 특징(제1 모달리티 특징)에 완전 연결 레이어를 통과한 특징(제2 모달리티 특징) f_{vid1}, f_{EEG1} 을 융합하여 최종적인 융합 특징 $f_{fusion2}$를 추출하고, 이를 이용하여 모달리티의 어텐션을 계산할 수 있다.

[0040] 감정 인식 시스템은 하위 계층 분해(low-rank decomposition) 기반 쌍선형 풀링(bilinear pooling)을 이용하여 융합을 적용할 수 있다. 첫번째 레이어는 쌍선형 모델(bilinear model) 과정에 해당하고, 두번째 레이어는 삼선형(trilinear model)을 확장한 것이다. 최종적으로, 감정 인식 시스템은 비디오 특징 및 EEG 특징을 이용하여 융합된 발란스 정보를 수학식 1과 같이 각 네트워크에서 출력된 정보를 어텐션 가중치로 가중합(weight sum)하여 출력할 수 있다. $\bar{\alpha} = [\bar{\alpha}_{vid}, \bar{\alpha}_{EEG}]$ 는 $\hat{\alpha} = [\hat{\alpha}_{vid}, \hat{\alpha}_{EEG}]$의 합을 1로 정규화한 결과이다.

[0041] 수학식 1: $\hat{y}_{fusion} = \bar{\alpha}_{vid}\,\hat{y}_{img} + \bar{\alpha}_{EEG}\hat{y}_{EEG}$

[0042] 멀티모달 어텐션 네트워크에서는 각 모달리티 특징(제1 모달리티 특징) f_{vid}, f_{EEG}을 융합하는 과정에서 하위 계층 분해(low-rank decomposition) 과정을 이용하여 기존의 쌍선형 풀링 과정에서 필요한 방대한 가중치 파라미터와 메모리 사이즈를 감소시키고, 과적합(overfitting)을 감소시킬 수 있다. 실시예에서 제안하는 융합 과정은 다음과 같다. 먼저, 비디오 특징과 EEG 특징을 외적(outer product)하면 수학식 2와 같이 나타낼 수 있다.

[0043] 수학식 2: $F = f_{vid} \otimes f_{EEG}$

[0044] 한편, 텐서(tensor) 형태로 구성되어 있는 가중치 $\mathcal{W}$에 대하여 하위 계층 분해를 적용하면 수학식 3과 같이 근사화될 수 있다. 이때, r은 텐서의 랭크(rank)를 의미할 수 있다.

[0045] 수학식 3:

[0046] $$W \simeq \sum_{i=1}^{r} w_{vid}^{(i)} \otimes w_{vid}^{(i)}$$

[0047] 특징 F와 근사화된 가중치 $\mathcal{W}$를 적용함에 따라 융합된 특징(제1 융합 특징) $f_{fusion1}$은 수학식 4와 같이 표현할

수 있다.

[0048] 수학식 4:

[0049] $$f_{fusion1} = \left(\sum_{i=1}^{r} w_{vid}^{(i)} \otimes w_{EEG}^{(i)}\right) \cdot (f_{vid} \otimes f_{EEG}) = \left(\sum_{i=1}^{r} w_{vid}^{(i)} \cdot f_{img}\right) \circ \left(\sum_{i=1}^{r} w_{EEG}^{(i)} \cdot f_{EEG}\right)$$

[0050] 수학식 4에서 연산자 $\circ$ 는 다중 곱(Elementwise Multiplication)이다. 최종적으로 분해(decomposition)된 가중치들은 각 모달리티 특징 벡터에 적용되고 다중곱을 통해 융합될 수 있다. 하위 계층 분해를 적용한 쌍선형 풀링은 1차원의 벡터와 2차원의 매트릭스에 대한 멀티곱을 반복적으로 수행하기 때문에 2차원의 매트릭스와 3차원의 텐서의 멀티곱의 연산으로 이루어져있는 기존의 쌍선형 모델에 비하여 연산량이 크게 감소하고, 가중 파라미터 또한 감소되면서 근사화될 수 있다.

[0051] 초기의 융합 특징(제1 융합 특징) $f_{fusion1}$ 과 각 모달리티 특징(제2 모달리티 특징) f_{vid1}, f_{EEG1} 을 융합하는 삼선형 모델을 하위 랭크 분해로 근사화한 식은 수학식 5와 같다.

[0052] 수학식 5:

[0053] $$f_{fusion2} = \left(\sum_{i=1}^{r} w_{vid1}^{(i)} \cdot f_{vid1}\right) \circ \left(\sum_{i=1}^{r} w_{EEG1}^{(i)} \cdot f_{EEG1}\right) \circ \left(\sum_{i=1}^{r} w_{fusion1}^{(i)} \cdot f_{fusion1}\right)$$

[0054] 예를 들면, 랭크가 8로 설정될 수 있고, f_{vid}, f_{EEG} 는 2048 및 3400 길이의 벡터이며, $w_{vid}^{(i)}$를 $w_{EEG}^{(i)}$를 통해 512 길이의 특징 벡터로 투영(projection)하도록 할 수 있다. 또한, $f_{fusion1}$ 및 $f_{fusion2}$의 길이는 256으로 설정될 수 있다.

[0055] 감정 인식 시스템은 멀티모달 어텐션 네트워크를 학습시킬 수 있다. 비디오 모달리티 네트워크 및 EEG 모달리티 네트워크는 각각의 학습 과정을 수행하여 프리즈(freeze)할 수 있다. 감정 인식 시스템은 입력 정보에 대하여 비디오 모달리티 네트워크 및 EEG 모달리티 네트워크에 대한 출력 결과를 확인하고, 실측 자료(ground truth)와 에러를 측정할 수 있다. 감정 인식 시스템은 비디오 모달리티 네트워크 출력의 에러가 작은 경우, 어텐션 가중치의 라벨을 α=[1, 0]으로 정의하고 EEG 모달리티의 에러가 작은 경우, α=[1, 0]으로 정의할 수 있다.

[0056] 감정 인식 시스템은 입력(비디오 데이터, EEG 신호)에 대하여 비디오 모달리티 네트워크 및 EEG 모달리티 네트워크에서 추출된 잠정적인(interim) 특징 f_{vid}, f_{EEG} 를 어텐션 네트워크의 입력으로 사용할 수 있다. 감정 인식 시스템은 어텐션 네트워크에서 출력된 어텐션 가중치 $\hat{\alpha}$ 를 어텐션 라벨 α와 오차를 측정하여 트레이닝 손실(training loss)로 정의할 수 있다. 감정 인식 시스템은 트레이닝 손실 계산 시 어텐션 가중치에 대한 MSE 손실만 이용하지 않는다. 각 모달리티에 auxiliary 레이어 FC_2를 이용하여 감정 출력 값을 출력하고, 이에 대한 오차를 트레이닝 손실에 추가할 수 있다.

[0057] 수학식 6:

[0058] $$L_{fusion} = L_{MSE}(\alpha, \hat{\alpha}) + L_{MSE}(y, \hat{y}'_{vid}) + L_{MSE}(y, \hat{y}'_{EEG})$$

[0059] 두번째 융합 레이어에 입력으로 이용되는 특징 f_{vid1}, f_{EEG1} 를 개선(refinement)하는 역할을 수행함으로써 최종적인 융합 특징(제2 융합 특징) 역시 개선될 수 있다. 추론 과정에서는 training/inference와 관련된 path에 대하여 연산을 수행하고 auxiliary 레이어 FC_2(training only와 관련된 path)에 대해서는 추론을 수행하지 않는다.

[0060] 모달리티 어텐션 네트워크를 학습하는 과정에서 모달리티 어텐션을 바이너리로 선택하는 분류 태스크(classification task)로 학습하지 않았고, 가중치 값을 회귀하기 위한 MSE로 손실을 계산할 수 있다. 이는, 각 모달리티에 대한 어텐션 가중치가 바이너리가 아닌 형태로 출력하기 위해서이다.

[0061] 도 3은 일 실시예에 따른 감정 인식 시스템에서 긍정/부정을 추정한 결과를 나타낸 예이다.

[0062] 감정 인식 시스템에서 발란스(긍정/부정)를 추정한 결과를 나타낸 예이다. 실시예에서 제안된 융합 결과는 영상 또는 EEG의 각 단일 모달리티 대비 실측 자료에 근접해지는 것을 확인할 수 있다.

[0063] 또한, MAHNOB_HCI dataset에서 모달리티 및 융합 방식에 따른 Pearson Correlation Coefficient(PCC)및 root mean square error(RMSE)를 비교한 것이다. 표 1은 비특허문헌 <M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. "Analysis of EEG signals and facial expressions for continuous emotion detection," IEEE Transactions on Affective Computing, vol. 7 no. 1, pp. 17-28, Mar. 2016>과 실시예에서 제안된 방법을 비교한 것이다. 여기서, PCC는 높을 수록 우수한 성능이며, RMSE는 낮을 수록 우수한 성능을 의미한다.

[0064] 표 1:

	Soleymani et al.'s [5]*		Proposed	
	PCC	RMSE	PCC	RMSE
Image modality only	0.48±0.37	0.0430±0.0260	0.52±0.07	0.0393±0.0044
EEG modality only	0.24±0.37	0.0530±0.0290	0.29±0.11	0.0485±0.0047
Feature level fusion (FDF)	0.40±0.33	0.0470±0.0250	-	-
Decision level fusion (DLF)	0.45±0.35	0.0440±0.0260	0.52±0.09	0.0373±0.0034
Multimodal attention network	-	-	**0.53**±0.08	**0.0366**±0.0032

[0065]

[0066] 일 실시예에 따른 감정 인식 시스템은 시간 영역에서 연속성있는 어노케이션을 라벨로 이용하고 시스템 인과관계를 보정하면서 실시간으로 발란스 값을 출력하도록 설계될 수 있다. 인간의 실제 감정반응을 모니터링한 MAHNOB-HCI dataset에서 제안하는 융합 방법을 통하여 비디오 단일 모달리티 대비 약 6.9%의 성능 개선을 보이는 것을 확인할 수 있다.

[0067] 도 4는 일 실시예에 따른 감정 인식 시스템의 구성을 설명하기 위한 블록도이고, 도 5는 일 실시예에 따른 감정 인식 시스템에서 감정 인식 방법을 설명하기 위한 흐름도이다.

[0068] 감정 인식 시스템(100)에 포함된 프로세서는 추출부(410), 결정부(420) 및 출력부(430)를 포함할 수 있다. 이러한 프로세서 및 프로세서의 구성요소들은 도 5의 감정 인식 방법이 포함하는 단계들(510 내지 530)을 수행하도록 감정 인식 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서의 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다. 여기서, 프로세서의 구성요소들은 감정 인식 시스템에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다.

[0069] 프로세서는 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 감정 인식 시스템(100)에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 감정 인식 시스템을 제어할 수 있다.

[0070] 단계(510)에서 추출부(410)는 비디오 신호와 EEG 신호로부터 각각의 모달리티 특징을 추출할 수 있다. 추출부(410)는 입력 영상 시퀀스에서 크롭(crop)된 얼굴 이미지를 획득하고, 획득된 얼굴 이미지를 딥 컨볼루션 인코더(deep convolution encoder)를 이용하여 비디오 모달리티 특징을 추출하고, 추출된 비디오 모달리티 특징을 비디오 모달리티 네트워크를 통해 회귀(regression)를 수행할 수 있다. 추출부(410)는 획득된 얼굴 이미지를 CNN 기반의 딥 컨볼루션 인코더를 이용하여 비디오 신호를 1차원의 특징 벡터로 변환하고, 변환된 1차원의 특징 벡터를 비디오 모달리티 네트워크를 통과시킴에 따라 출력되는 히든 스테이트 벡터(hidden state vector)를 완전 연결 레이어에 통과시켜 입력 영상 시퀀스의 발란스 라벨 값을 출력하는 단계를 포함할 수 있다. 이때, 비디오 모달리티 네트워크는, LSTM 네트워크를 포함하고, 변환된 1차원의 특징 벡터를 상기 LSTM 네트워크를 통과시키기 전에 완전 연결 레이어를 통과시켜 특징 벡터의 차원이 조정될 수 있다. 추출부(410)는 EEG 신호로부터 시간 영역, 주파수 영역 및 시간-주파수 영역에 대한 EEG 모달리티 특징을 추출하고, 추출된 EEG 모달리티 특징을 EEG 모달리티 네트워크에 입력하여 EEG 신호의 발란스 라벨 값을 출력할 수 있다. 이때, EEG 모달리티 네트워크는, LSTM 네트워크를 포함할 수 있다.

[0071] 단계(520)에서 결정부(420)는 추출된 각각의 모달리티 특징을 멀티모달 어텐션 네트워크(Multimodal attention network)에 입력하여 모달리티에 대한 어텐션 가중치를 결정할 수 있다. 결정부(420)는 비디오 신호로부터 추출된 제1 비디오 모달리티 특징과 EEG 신호로부터 추출된 제1 EEG 모달리티 특징을 융합하여 제1 융합 특징을

생성하고, 제1 비디오 모달리티 특징 및 제1 EEG 모달리티 특징을 멀티모달 어텐션 네트워크의 완전 연결 레이어를 통과시킴에 따라 제2 비디오 모달리티 특징 및 제2 EEG 모달리티 특징을 획득할 수 있다. 결정부(420)는 생성된 제1 융합 특징과 제2 비디오 모달리티 특징 및 제2 EEG 모달리티 특징을 융합하여 제2 융합 특징을 획득하고, 획득된 제2 융합 특징을 이용하여 각각의 모달리티의 어텐션을 계산할 수 있다. 이때, 멀티모달 어텐션 네트워크는, 하위 계층 분해(low-rank decomposition) 기반 쌍선형 풀링(bilinear pooling)을 멀티-레이어(multi-layer)로 구성되고, 각각의 모달리티 특징을 융합시킴에 따라 획득된 발란스 정보는, 멀티모달 어텐션 네트워크를 통해 각각 출력된 정보가 어텐션 가중치로 가중합(weight sum)되어 출력될 수 있다. 결정부(420)는 추출된 제1 비디오 모달리티 특징과 추출된 제1 EEG 모달리티 특징을 하위 계층 분해를 적용함에 따라 획득된 가중치들을 각각의 모달리티 특징 벡터에 적용시켜 멀티곱을 통해 융합시킬 수 있다. 결정부(420)는 제1 융합 특징과 상기 제2 비디오 모달리티 특징 및 제2 EEG 모달리티 특징을 하위 계층 분해(low-rank decomposition)에 기반하여 융합시킬 수 있다. 결정부(420)는 비디오 신호를 비디오 모달리티 네트워크를 통해 추출된 제1 비디오 모달리티 특징, EEG 신호를 EEG 모달리티 네트워크를 통해 추출된 제1 EEG 모달리티 특징을 멀티모달 어텐션 네트워크에 입력함에 따라 출력된 어텐션 가중치를 어텐션 라벨과 오차를 측정하여 트레이닝 손실을 계산할 수 있다.

[0072] 단계(530)에서 출력부(430)는 결정된 어텐션 가중치를 추출된 각각의 모달리티 특징에 반영하여 융합된 감정 정보를 출력할 수 있다.

[0073] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0074] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0075] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

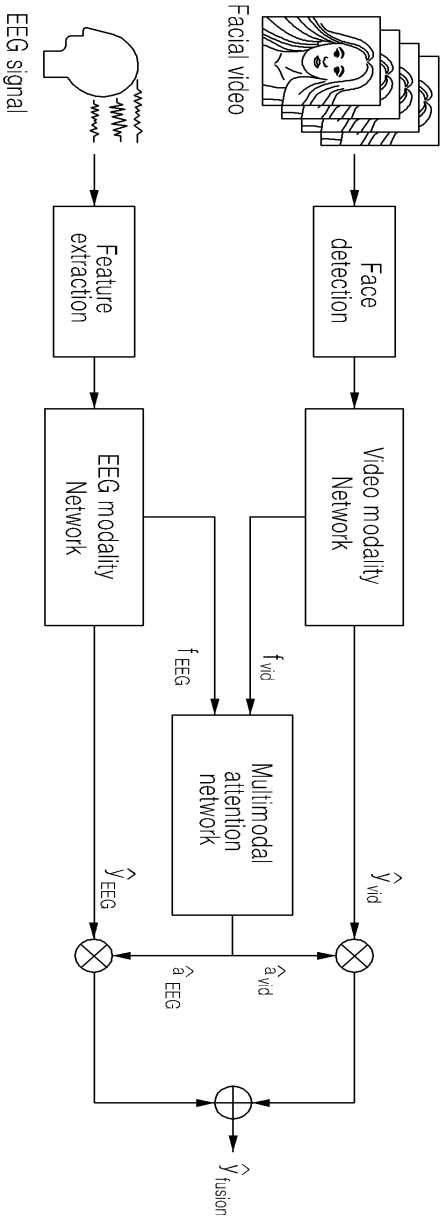
[0076] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될

수 있다.

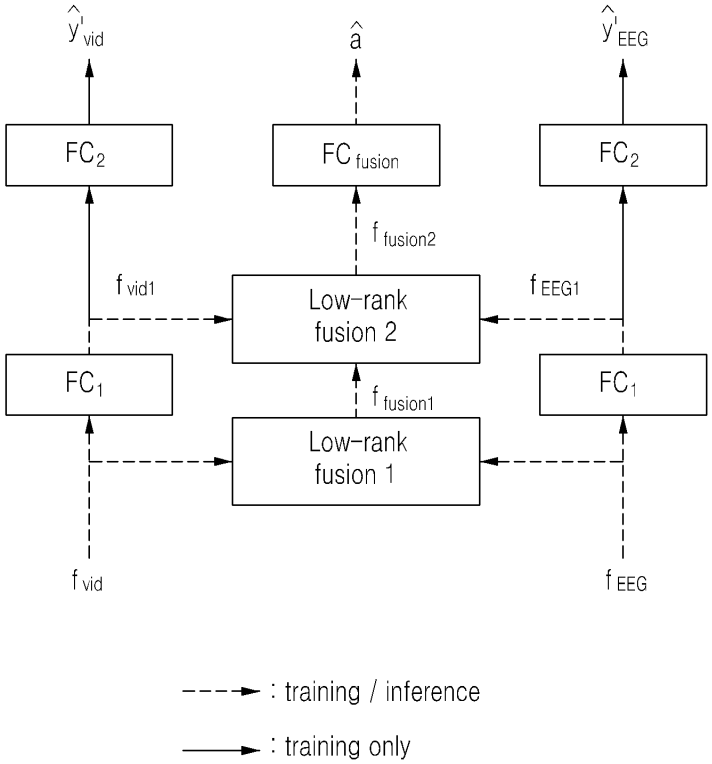
[0077] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

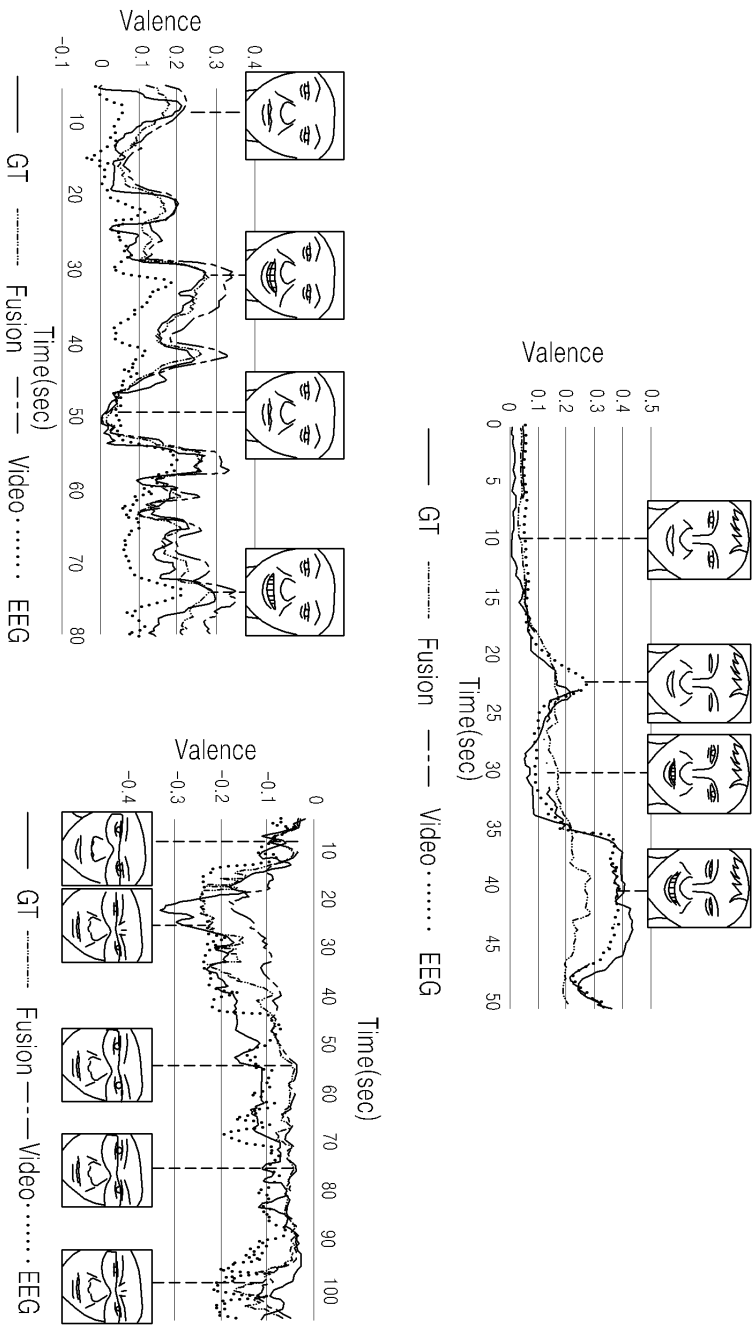
도면1



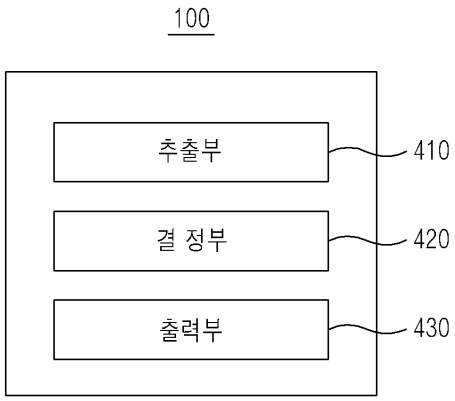
도면2


$\hat{y}'_{vid}$
$\hat{a}$
$\hat{y}'_{EEG}$
FC_2
FC_{fusion}
FC_2
$f_{fusion2}$
f_{vid1}
Low-rank fusion 2
f_{EEG1}
FC_1
$f_{fusion1}$
FC_1
Low-rank fusion 1
f_{vid}
f_{EEG}
: training / inference
: training only

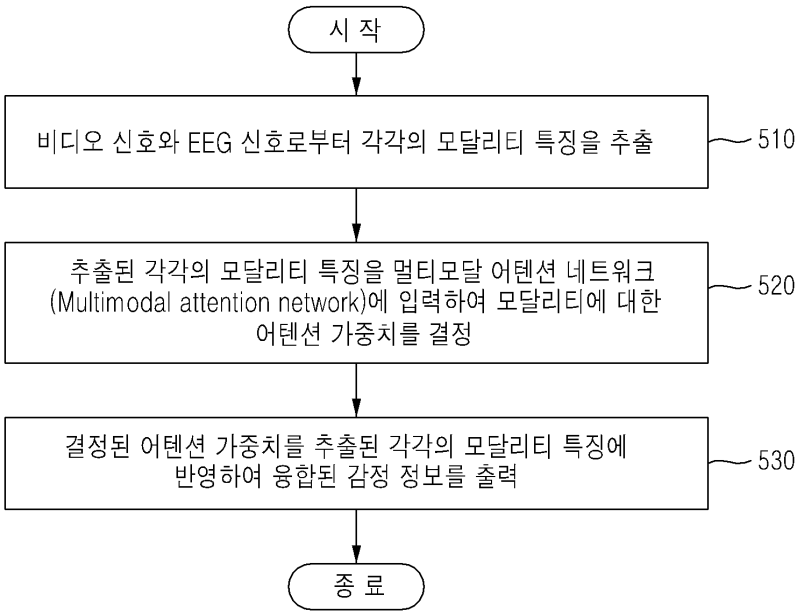
도면3



도면4


100
추출부
410
결 정부
420
출력부
430

도면5


시 작
비디오 신호와 EEG 신호로부터 각각의 모달리티 특징을 추출
510
추출된 각각의 모달리티 특징을 멀티모달 어텐션 네트워크 (Multimodal attention network)에 입력하여 모달리티에 대한 어텐션 가중치를 결정
520
결정된 어텐션 가중치를 추출된 각각의 모달리티 특징에 반영하여 융합된 감정 정보를 출력
530
종 료