(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0128933
(43) 공개일자 2019년11월19일

(51) 국제특허분류(Int. Cl.)
A61B 5/16 (2006.01) *A61B 5/00* (2006.01)
(52) CPC특허분류
A61B 5/165 (2013.01)
A61B 5/7246 (2013.01)
(21) 출원번호 10-2018-0053306
(22) 출원일자 2018년05월09일
심사청구일자 2018년05월09일

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

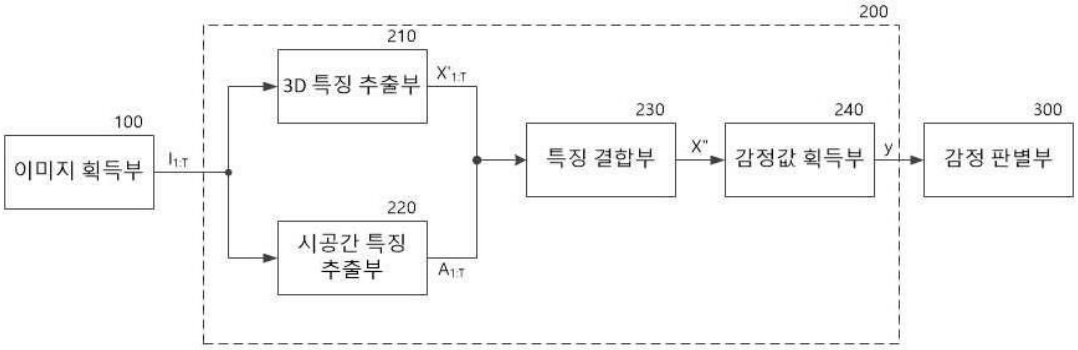
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **시공간 주의 기반 감정 인식 장치 및 방법**

(57) 요 약

감정 인식 장치 및 방법을 공개한다. 본 발명의 감정 인식 장치 및 방법은 다수의 프레임을 포함하는 이미지 시퀀스로부터 3차원 특징을 획득함과 동시에 각 프레임에 대한 시공간 특징을 추출하여 시공간 가중치로 획득하고, 3차원 특징에 시공간 가중치를 가중함으로써, 별도의 관심 영역을 설정하지 않더라도 정확한 감정을 판별할 수 있다.

대 표 도



이 발명을 지원한 국가연구개발사업
과제고유번호 R0124-16-0002
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터
연구사업명 첨단융복합콘텐츠기술개발사업
연구과제명 상대방의 감성을 추론, 판단하여 그에 맞추어 대화하고 대응할 수 있는 감성 지능 연구개발
기 여 율 1/1
주관기관 한국과학기술원
연구기간 2017.09.01 ~ 2018.06.30

명 세 서

청구범위

청구항 1

기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 이미지 시퀀스의 시간적으로 연속하는 T개(여기서 T는 자연수)의 프레임을 3차원의 단일 이미지로서 패턴 인식하여 3D 특징을 추출하는 3D 특징 추출부;

기지정된 2차원 패턴 인식 기법에 따라 미리 학습되어, 상기 T개의 프레임 각각으로부터 패턴 인식을 통해 T개의 공간적 특징을 추출하고, 획득된 T개의 공간적 특징 사이의 시공간 특징을 추가하여 시공간 가중치를 획득하는 시공간 특징 추출부;

상기 시공간 가중치를 상기 3D 특징에 가중하여 감정 특징을 획득하고, 기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 상기 감정 특징으로부터 미리 지정된 범위 이내의 값을 갖는 감정값을 추출하는 감정값 추출부; 및

감정값에 대비한 감정이 미리 저장되어, 상기 감정값 추출부에서 획득된 상기 감정값에 대응하는 감정을 판별하는 감정 판별부; 를 포함하는 감정 인식 장치.

청구항 2

제1 항에 있어서, 상기 3D 특징 추출부는

미리 학습된 3D CNN(3D Convolutional Neural Networks)을 포함하여, 상기 3D 특징을 추출하는 감정 인식 장치.

청구항 3

제1 항에 있어서, 상기 시공간 특징 추출부는

미리 학습된 2D CNN(2D Convolutional Neural Networks)을 포함하여, 상기 T개의 프레임 각각에 대한 상기 T개의 공간적 특징을 추출하는 공간 인코더;

미리 학습된 ConvLSTM(Convolutional Long Short-Term Memory)을 포함하여, 상기 T개의 공간적 특징 사이의 시공간 특징을 추출하는 시간 디코더; 및

상기 시간 디코더에서 추출된 시공간 특징을 기지정된 방식으로 정규화하여, 상기 시공간 가중치를 획득하는 정규화부; 를 포함하는 감정 인식 장치.

청구항 4

제3 항에 있어서, 상기 공간 인코더는

상기 2D CNN가 각각 다수의 필터를 포함하는 컨볼루션 레이어, ReLU(Rectified Linear Unit) 레이어 및 맥스 풀링(Max-Pooling) 레이어를 포함하여 상기 공간적 특징의 공간 해상도를 상기 프레임의 공간 해상도보다 낮도록 축소하는 감정 인식 장치.

청구항 5

제4 항에 있어서, 상기 시간 디코더는

상기 ConvLSTM가 다수의 ConvLSTM 레이어를 포함하여, 순차적 디콘볼루션을 수행함으로써, 상기 공간적 특징의 축소된 공간 해상도를 복구하는 감정 인식 장치.

청구항 6

제5 항에 있어서, 상기 정규화기는

소프트 맥스 함수를 이용하여, 상기 시공간 특징을 정규화하는 감정 인식 장치.

청구항 7

제6 항에 있어서, 상기 감정값 추출부는

상기 3D 특징과 상기 시공간 가중치를 하다마드 곱셈하여 상기 감정 특징을 획득하는 특징 결합부; 및

미리 학습된 3D CNN을 포함하여 상기 감정 특징으로부터 감정을 대표하는 감정값을 추출하는 감정값 획득부; 를 포함하는 감정 인식 장치.

청구항 8

제7 항에 있어서, 상기 감정값 획득부는

상기 감정값을 기지정된 범위 이내의 스칼라 값으로 획득하는 감정 인식 장치.

청구항 9

제7 항에 있어서, 상기 감정값은

감정을 각성(Arousal) 및 유인가(Valence)를 2개의 축으로 하여 2차원으로 표현하는 감정 모델에서 상기 유인가의 값을 나타내는 감정 인식 장치.

청구항 10

제1 항에 있어서, 상기 감정 인식 장치는

다수의 프레임을 포함하는 이미지 시퀀스에서 시간적으로 연속하는 상기 T개의 프레임을 분리하여 출력하는 이미지 획득부; 를 더 포함하는 감정 인식 장치.

청구항 11

기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 이미지 시퀀스의 시간적으로 연속하는 T개(여기서 T는 자연수)의 프레임을 3차원의 단일 이미지로서 패턴 인식하여 3D 특징을 추출하는 단계;

기지정된 2차원 패턴 인식 기법에 따라 미리 학습되어, 상기 T개의 프레임 각각으로부터 패턴 인식을 통해 T개의 공간적 특징을 추출하고, 획득된 T개의 공간적 특징 사이의 시공간 특징을 추가하여 시공간 가중치를 획득하는 단계;

상기 시공간 가중치를 상기 3D 특징에 가중하여 감정 특징을 획득하는 단계;

기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 상기 감정 특징으로부터 미리 지정된 범위 이내의 값을 갖는 감정값을 추출하는 단계; 및

상기 감정값에 대응하는 감정을 판별하는 단계; 를 포함하는 감정 인식 방법.

청구항 12

제11 항에 있어서, 상기 3D 특징을 추출하는 단계는

미리 학습된 3D CNN(3D Convolutional Neural Networks)을 이용하여 상기 3D 특징을 추출하는 감정 인식 방법.

청구항 13

제11 항에 있어서, 상기 시공간 가중치를 획득하는 단계는

미리 학습된 2D CNN(2D Convolutional Neural Networks)을 이용하여 상기 T개의 프레임 각각에 대한 상기 T개의 공간적 특징을 추출하는 단계;

미리 학습된 ConvLSTM(Convolutional Long Short-Term Memory)을 이용하여 상기 T개의 공간적 특징 사이의 시공간 특징을 추출하는 단계; 및

추출된 시공간 특징을 기지정된 방식으로 정규화하여, 상기 시공간 가중치를 획득하는 단계; 를 포함하는 감정 인식 방법.

청구항 14

제11 항에 있어서, 상기 감정값을 추출하는 단계는

미리 학습된 3D CNN을 포함하여 상기 감정 특징으로부터 감정을 대표하는 감정값을 추출하는 감정 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 감정 인식 장치 및 방법에 관한 것으로, 특히 시공간 주의 기반 감정 인식 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 감정 인식은 대화형 시스템에서 중요한 이슈 중 하나이다. 대화형 시스템은 기존의 명령 입력 방식이 아닌 사용자와의 상호 대화를 통해 사용자의 요구 사항을 판별한다. 이때, 감정 인식 기술이 적용되면, 사용자의 요구 사항을 더욱 정확하게 판별할 수 있다는 장점이 있다.

[0003] 또한 감정 인식은 통증이나 심리적 고통 탐지와 같이 의료 분야 등에 적용될 수 있으며 그 외에도 다양한 분야에 적용될 수 있다.

[0004] 기존의 감정 인식에 대한 연구는 대부분 감정을 공포, 분노, 행복, 혐오, 슬픔, 놀람과 같은 기지정된 개수(예를 들면 6가지)로 지정된 기본 감정에 따라 이산된 범주로 분류하는 범주형 감정 인식 방식이 대부분이었다. 그러나 범주형 감정 인식은 지정된 감정으로만 분류하여 인식함에 따라 분류되지 않는 감정의 영역이 존재될 뿐만 아니라, 인식 가능한 감정의 종류가 제한되는 한계가 있다.

[0005] 도1 은 사람의 감정을 나타내는 이미지의 일예를 나타낸다.

[0006] 도1 은 (a) 내지 (d)는 Ekman이 정의한 4 가지 유형의 놀람에 대한 표정 이미지로서, (a)는 놀랄만한 질문(questioning surprise), (b)는 깜짝 놀람(astonished surprise), (c)는 어리둥절한 놀람(dazed surprise)을 나타내고, (d)는 완전히 놀람(full surprise)을 표현하고 있다.

[0007] 도1 에 도시된 바와 같이, 놀람에도 다양한 놀람이 존재할 수 있으나, 기존의 범주형 감정 인식은 모두 놀람으로만 분류될 뿐, 미묘한 감정의 차이를 인식할 수 없다는 한계가 있다. 이에 감정을 연속되는 2개의 영역(domain)에 따라 2차원으로 표현하는 방안이 제안되었다.

[0008] 도2 는 연속하는 2차원 그래프로 나타나는 감정의 일예를 나타낸다.

[0009] 도2 에서 2차원의 각 축은 각성(Arousal) 및 유인가(Valence)를 나타내고, 각성축은 활동적인지 비활동적의 수준을 나타내고, 유인가축은 긍정적 또는 부정적인 수준을 나타낸다. 도2 에서 도시된 바와 같이, 연속되는 2차원으로 감정을 묘사하는 방식은 기존 범주형에 비해 더 복잡하고 미묘한 감정을 표현할 수 있다.

[0010] 한편 최근에는 감정 인식 기법에 신경망(neural network)를 적용하여 감정 인식의 정확도를 향상시키고 있다. 그러나 기존의 감정 인식 기법은 대부분 단일 이미지로부터 감정을 인식하도록 연구가 수행되어, 시간에 따른 이미지 시퀀스(image sequence)로부터 사람의 감정을 정확하게 인식하는 방법에 대한 연구가 부족한 실정이다. 실제 사람의 감정은 시간의 흐름에 따라 서서히 연속되어 변화되므로, 연속되는 이미지 시퀀스를 이용하여 감정을 인식하는 경우, 단일 이미지보다 더욱 정확하게 감정을 인식할 수 있다.

[0011] 또한 기존의 감정 인식은 도1 에 도시된 바와 같이, 사람의 얼굴 이미지에서 감정 표출이 강하게 나타나는 것으로 예상되는 관심 영역을 미리 지정하고, 지정된 관심 영역에 대해 분석을 수행한다. 그러나 일부의 관심 영역만을 활성화하여 표정을 추정하고, 감정을 인식함에 따라 다양한 얼굴 이미지에 대해 최적의 성능으로 감정을 인식할 수 없다는 한계가 있다.

선행기술문헌

특허문헌

[0012] (특허문헌 0001) 한국 공개 특허 제10-2013-0015958호 (2013.02.14 공개)

발명의 내용

해결하려는 과제

[0013] 본 발명의 목적은 얼굴 이미지 시퀀스로부터 시간적, 공간적 주의에 기반하여 감정을 인식할 수 있는 감정 인식 장치 및 방법을 제공하는데 있다.

[0014] 본 발명의 다른 목적은 얼굴 이미지 시퀀스에 관심 영역을 지정하지 않고도 감정을 인식할 수 있는 감정 인식 장치 및 방법을 제공하는데 있다.

[0015] 본 발명의 또 다른 목적은 얼굴 이미지 시퀀스로부터 2차원 및 3차원 특징을 각각 추출하고, 추출된 2차원 및 3차원 특징을 이용하여 감정을 정확하게 인식할 수 있는 감정 인식 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0016] 상기 목적을 달성하기 위한 본 발명의 일 예에 따른 감정 인식 장치는 기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 이미지 시퀀스의 시간적으로 연속하는 T개(여기서 T는 자연수)의 프레임을 3차원의 단일 이미지로서 패턴 인식하여 3D 특징을 추출하는 3D 특징 추출부; 기지정된 2차원 패턴 인식 기법에 따라 미리 학습되어, 상기 T개의 프레임 각각으로부터 패턴 인식을 통해 T개의 공간적 특징을 추출하고, 획득된 T개의 공간적 특징 사이의 시공간 특징을 추가하여 시공간 가중치를 획득하는 시공간 특징 추출부; 상기 시공간 가중치를 상기 3D 특징에 가중하여 감정 특징을 획득하고, 기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 상기 감정 특징으로부터 미리 지정된 범위 이내의 값을 갖는 감정값을 추출하는 감정값 추출부; 및 감정값에 대비한 감정이 미리 저장되어, 상기 감정값 추출부에서 획득된 상기 감정값에 대응하는 감정을 판별하는 감정 판별부를 포함한다.

[0017] 상기 3D 특징 추출부는 미리 학습된 3D CNN(3D Convolutional Neural Networks)을 포함하여, 상기 3D 특징을 추출할 수 있다.

[0018] 상기 시공간 특징 추출부는 미리 학습된 2D CNN(2D Convolutional Neural Networks)을 포함하여, 상기 T개의 프레임 각각에 대한 상기 T개의 공간적 특징을 추출하는 공간 인코더; 미리 학습된 ConvLSTM(Convolutional Long Short-Term Memory)을 포함하여, 상기 T개의 공간적 특징 사이의 시공간 특징을 추출하는 시간 디코더; 및 상기 시간 디코더에서 추출된 시공간 특징을 기지정된 방식으로 정규화하여, 상기 시공간 가중치를 획득하는 정규화부를 포함할 수 있다.

[0019] 상기 공간 인코더는 상기 2D CNN가 각각 다수의 필터를 포함하는 컨볼루션 레이어, ReLU(Rectified Linear Unit) 레이어 및 맥스 풀링(Max-Pooling) 레이어를 포함하여 상기 공간적 특징의 공간 해상도를 상기 프레임의 공간 해상도보다 낮도록 축소할 수 있다.

[0020] 상기 시간 디코더는 상기 ConvLSTM가 다수의 ConvLSTM 레이어를 포함하여, 순차적 디콘볼루션을 수행함으로써, 상기 공간적 특징의 축소된 공간 해상도를 복구할 수 있다.

[0021] 상기 정규화기는 소프트 맥스 함수를 이용하여, 상기 시공간 특징을 정규화할 수 있다.

[0022] 상기 감정값 추출부는 상기 3D 특징과 상기 시공간 가중치를 하다마드 곱셈하여 상기 감정 특징을 획득하는 특징 결합부; 및 미리 학습된 3D CNN을 포함하여 상기 감정 특징으로부터 감정을 대표하는 감정값을 추출하는 감정값 획득부를 포함할 수 있다.

[0023] 상기 목적을 달성하기 위한 본 발명의 다른 예에 따른 감정 인식 방법은 기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 이미지 시퀀스의 시간적으로 연속하는 T개(여기서 T는 자연수)의 프레임을 3차원의 단일 이미지로서 패턴 인식하여 3D 특징을 추출하는 단계; 기지정된 2차원 패턴 인식 기법에 따라 미리 학습되어, 상기 T개의 프레임 각각으로부터 패턴 인식을 통해 T개의 공간적 특징을 추출하고, 획득된 T개의 공간적 특징 사이의 시공간 특징을 추가하여 시공간 가중치를 획득하는 단계; 상기 시공간 가중치를 상기 3D 특징에 가중하여 감정 특징을 획득하는 단계; 기지정된 3차원 패턴 인식 기법에 따라 미리 학습되어, 상기 감정 특징으로부터 미리 지정된 범위 이내의 값을 갖는 감정값을 추출하는 단계; 및 상기 감정값에 대응하는 감정을 판별하는 단계를 포함한

다.

발명의 효과

[0024] 따라서, 본 발명의 감정 인식 장치 및 방법은 이미지 시퀀스로부터 2차원 및 3차원 특징을 각각 획득하고, 획득된 2차원 및 3차원 특징을 함께 이용하여 정확하게 감정을 인식할 수 있다. 또한 시간적 및 공간적 주의에 기반하여 감정을 인식할 뿐만 아니라, 감정을 인식하기 위한 영역을 별도로 지정하지 않고도 감정을 연속적인 유인가를 기반으로 정확하게 인식할 수 있다.

도면의 간단한 설명

[0025] 도1 은 사람의 감정을 나타내는 이미지의 일예를 나타낸다.

도2 는 연속하는 2차원 그래프로 나타나는 감정의 일예를 나타낸다.

도3 은 본 발명의 일 실시예에 따른 감정 인식 장치의 개략적 구성을 나타낸다.

도4 는 도3 의 시공간 특징 추출부의 상세 구성의 일예를 나타낸다.

도5 는 도3 의 감정 인식 장치의 학습 방법을 설명하기 위한 도면이다.

도6 은 본 발명의 일 실시예에 따른 감정 인식 방법을 나타낸다.

도7 은 본 실시예의 시공간 가중치를 시각화한 도면이다.

도8 및 도9 는 각각 2 종류의 RECOLA 데이터 세트와 AV + EC 데이터 세트에 대해 본 실시예에 따른 감정 인식 방법을 적용하여 획득되는 감정값과 검증값을 비교한 결과를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0026] 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.

[0027] 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.

[0028] 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.

[0029] 도3 은 본 발명의 일 실시예에 따른 감정 인식 장치의 개략적 구성을 나타내고, 도4 는 도3 의 시공간 특징 추출부의 상세 구성의 일예를 나타낸다.

[0030] 도3 을 참조하면, 본 실시예에 따른 감정 인식 장치는 감정을 인식해야 하는 대상이 포함된 이미지 시퀀스를 획득하는 이미지 획득부(100), 이미지 획득부(100)에서 전달된 이미지 시퀀스로부터 감정값을 추출하는 감정 추출부(200) 및 추출된 감정값에 따라 이미지에 포함된 대상의 감정을 판별하는 감정 판별부(300)를 포함한다.

[0031] 우선 이미지 획득부(100)는 감정을 인식해야 하는 대상이 포함된 이미지 시퀀스를 획득한다. 특히 본 실시예에서 이미지 획득부(100)는 단일 이미지가 아닌, 연속된 T(여기서 T는 자연수)개의 프레임(I_f)(여기서 f는 프레임 인덱스로서 자연수)을 포함하는 이미지 시퀀스($I_{1:T} = \{I_1, I_2, \cdots, I_T\}$)를 획득한다. 여기서 이미지 시퀀스($I_{1:T}$)의 각 프레임에는 감정을 인식할 수 있도록 대상의 얼굴이 포함된다.

[0032] 그리고 이미지 획득부(100)는 획득된 이미지 시퀀스를 감정 추출부(200)로 전달한다. 이때 이미지 획득부(100)는 획득된 이미지 시퀀스에 포함된 프레임의 개수가 T개를 초과하는 경우, 이미지 시퀀스에서 대상의 감정을 인식하고자 하는 시점의 프레임이 포함된 T개의 프레임을 분리하여 감정 추출부(200)로 전달할 수 있다. 예를 들면 이미지 획득부(100)는 100개의 프레임을 포함하는 이미지 시퀀스에서 제11 내지 제20 프레임(I_{11} ~ I_{20})를

별도로 분리(T가 10 인 것으로 가정)하여, 감정 추출부(200)로 전달할 수 있다.

[0033] 또한 이미지 시퀀스로부터 대상의 연속적인 감정 변화를 인식하고자 하는 경우에는, 이미지 시퀀스에서 순차적으로 T개의 프레임을 분리하여, 감정 추출부(200)로 전달할 수 있다. 일예로 제1 내지 제10 프레임(I_1 ~ I_{10})을 전달하고, 이후 제2 내지 제11 프레임(I_2 ~ I_{11})을 전달할 수 있다.

[0034] 감정 추출부(200)는 이미지 획득부(100)에서 전달된 이미지 시퀀스($I_{1:T}$)로부터 감정값을 추출한다. 특히 본 실시예에서 감정 추출부(200)는 이미지 시퀀스($I_{1:T}$)에 대해 미리 학습된 2차원(2D) 및 3차원(3D) 패턴 인식 기법을 이용하여 이미지 시퀀스($I_{1:T}$)의 특징을 추출하고, 추출된 특징을 결합하여 감정값을 추출한다.

[0035] 도3 에 도시된 바와 같이, 감정 추출부(200)는 3D 특징 추출부(210), 시공간 특징 추출부(220), 특징 결합부(230) 및 감정값 획득부(240)를 포함할 수 있다.

[0036] 3D 특징 추출부(210)는 대상의 감정을 판별하기 위해 이미지 획득부(100)에서 전달된 2차원의 이미지 시퀀스($I_{1:T}$)의 프레임($\{I_1, I_2, \cdots, I_T\}$) 전체를 3차원의 단일 객체로서 패턴 인식하여 3D 특징($X'_{1:T}$)을 추출한다. 즉 시간의 흐름에 따라 누적된 다수의 2차원 프레임을 포함하는 이미지 시퀀스($I_{1:T}$)를 3차원 이미지로 인식하여, 3차원의 이미지 시퀀스($I_{1:T}$)를 미리 지정된 패턴 인식 기법에 따라 분석함으로써 3D 특징($X'_{1:T}$)을 추출한다. 본 실시예에서는 3D 특징 추출부(210)가 시간에 따라 연속하는 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)을 포함하는 이미지 시퀀스($I_{1:T}$)로부터 감정 인식을 위한 3D 특징($X'_{1:T}$)을 추출하므로, 단일 이미지로부터 감정 인식을 위한 2D 특징을 추출하는 방식에 비해, 상대적으로 정확한 특징을 추출할 수 있다. 즉 대상의 감정을 매우 정확하게 판별할 수 있도록 한다.

[0037] 3D 특징 추출부(210)는 일예로 미리 학습된 3차원 콘볼루션 신경망(3D Convolutional Neural Networks: 이하 3D CNN)으로 구현될 수 있다.

[0038] 시공간 특징 추출부(220)는 이미지 획득부(100)에서 전달된 T개의 2차원 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각으로부터 시공간 주의(Spatiotemporal Attention)에 기반하여 특징을 추출한다. 특히 본 실시예에서 시공간 특징 추출부(220)는 이미지 시퀀스($I_{1:T}$)의 시공간 주의 기반 특징을 추출함으로써, 이미지 시퀀스($I_{1:T}$)에 대해 별도의 관심 영역을 지정하지 않더라도 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)내의 각 영역별 중요도에 따른 가중치를 획득할 수 있도록 한다.

[0039] 즉 본 실시예에 따른 감정 인식 장치는, 도1 과 같이 각 프레임($\{I_1, I_2, \cdots, I_T\}$)에서 사람의 얼굴에서 감정이 강하게 표출되는 영역(눈, 입)을 별도로 지정하지 않더라도, 시공간 특징 추출부(220)가 각 프레임의 영역별 감정 표출의 중요도를 시공간 주의에 기반하여 특징으로 추출하고, 추출된 특징을 시공간 가중치($A_{1:T}$)로서 3D 특징($X'_{1:T}$)에 부가함으로써 최적의 감정 인식 성능을 제공할 수 있다.

[0040] 이를 위해, 시공간 특징 추출부(220)는 도4 와 같이 구성될 수 있다.

[0041] 도4 를 참조하면, 시공간 특징 추출부(220)는 공간 주의(Spatial Attention) 기반 특징을 추출하기 위한 공간 인코더(221), 시공간 주의(Spatiotemporal Attention) 기반 특징을 추출하기 위한 시간 디코더(223) 및 추출된 특징을 지정된 범위 이내의 가중치로 변환하는 정규화하는 정규화기(225)를 포함한다.

[0042] 공간 인코더(221)는 이미지 시퀀스($I_{1:T}$)의 T개의 2차원 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각에 대해 공간적 특징($X_{1:T}$)을 추출하여 출력한다. 공간 인코더(221)는 지정된 2차원 패턴 인식 기법에 의해 미리 학습되어, T개의 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각의 공간적 패턴을 인식함으로써, 2차원의 공간적 특징($X_{1:T}$)을 추출한다.

[0043] 공간 인코더(221)는 일예로 미리 학습된 2차원 콘볼루션 신경망(2D Convolutional Neural Networks: 이하 2D CNN)으로 구현될 수 있다. 2D CNN은 2차원의 이미지에서 특징을 추출하기 위해 주로 이용되는 인공 신경망의 하나이다.

[0044] 공간 인코더(221)는 이미지 획득부(100)로부터 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)을 순차적으로 인가받아 공간적 특징($\{X_1, X_2, \cdots, X_T\}$)을 순차적으로 출력하도록 구성될 수 있으나, 시간을 줄이기 위해 T개의 프레임($\{I_1, I_2,$

…, I_T})을 동시에 인가받아 특징을 추출할 수 있도록 병렬로 구성될 수도 있다. 공간 인코더(221)가 병렬로 구성되는 경우, 모든 공간 인코더는 가중치 및 바이어스 값이 동일하게 공유되는 사이어미즈(Siamese) 네트워크로 구성된다.

[0045] 또한 공간 인코더(221)가 학습되는 과정에서 T개의 프레임({I_1, I_2, …, I_T})이 순차적으로 인가되더라도, T개의 프레임({I_1, I_2, …, I_T})에 대한 공간적 특징({X_1, X_2, …, X_T})이 모두 출력되기 이전에는 공간 인코더(221)의 가중치 및 바이어스 값이 가변되지 않아야 하며, T개의 프레임({I_1, I_2, …, I_T})에 대한 공간적 특징({X_1, X_2, …, X_T})이 모두 출력된 이후, 공간 인코더(221)의 가중치 및 바이어스 값이 가변될 수 있다. 이는 본 실시예에 따른 감정 인식 장치가 T개의 프레임({I_1, I_2, …, I_T})을 포함하는 이미지 시퀀스($I_{1:T}$)를 감정 인식을 위한 단위로 처리하기 때문이다.

[0046] 한편, 본 실시예에서 2D CNN으로 구현되는 공간 인코더(221)는 일예로 연속되는 3 X 3 컨볼루션 레이어와 ReLU(Rectified Linear Unit) 레이어 및 2 X 2 스트라이드(stride)의 맥스 풀링(Max-Pooling) 레이어를 포함하도록 구성될 수 있다. 여기서 3 X 3 컨볼루션 레이어와 ReLU 레이어 및 맥스 풀링 레이어는 각각 기지정된 개수의 필터를 포함할 수 있다. 일예로, 3 X 3 컨볼루션 레이어는 32개의 필터를 포함할 수 있고, ReLU 레이어는 64개의 필터를 포함할 수 있으며, 맥스 풀링 레이어는 128개의 필터를 포함하도록 구성될 수 있다.

[0047] 공간 인코더(221)가 3 X 3 컨볼루션 레이어와 ReLU 레이어 및 맥스 풀링 레이어를 포함하는 것은 이미지 시퀀스($I_{1:T}$)로부터 공간적 특징($X_{1:T}$)을 추출할 때, 매개 변수의 수를 줄임으로써 오버 피팅(overfitting) 문제를 방지하기 위함이다.

[0048] 시간 디코더(223)는 공간 인코더(221)에서 획득된 공간적 특징($X_{1:T}$)에 대해 시공간 주의 기반 특징을 추출한다.

[0049] 공간 인코더(221)가 2D CNN으로 구현되는 경우, 이미지 시퀀스($I_{1:T}$)의 T개의 프레임({I_1, I_2, …, I_T}) 각각에서의 공간적 특징, 즉 영역별 특징을 추출할 수 있다. 그러나 공간 인코더(221)가 T개의 프레임({I_1, I_2, …, I_T})을 개별적으로 특징을 추출함으로써, 시간적으로 연속하는 T개의 프레임({I_1, I_2, …, I_T}) 사이의 시간적 특징이 반영되지 않는 한계가 있다.

[0050] 이에 본 실시예에서 시간 디코더(223)는 시공간 주의 기반 특징을 추출함으로써, 공간적 특징($X_{1:T}$)에 시간적 특징이 더 부가되도록 한다. 시간 디코더(223)는 지정된 패턴 인식 기법에 의해 미리 학습되어, 공간적 특징({X_1, X_2, …, X_T})에 포함된 공간 패턴 특징을 가능한 유지하면서, 공간적 특징({X_1, X_2, …, X_T}) 중 시간적으로 서로 인접한 공간적 특징 사이의 시간적 특징을 추가로 추출한다.

[0051] 시간 디코더(223)는 일예로 미리 학습된 ConvLSTM(Convolutional Long Short-Term Memory)으로 구현될 수 있다. ConvLSTM 또한 인공 신경망의 하나로서, 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 하였다.

[0052] 여기서 시간 디코더(223)가 시간적 특징을 반영할 수 있는 LSTM이 아닌 ConvLSTM을 이용하는 것은 공간 인코더(221)에서 획득된 공간적 특징($X_{1:T}$)을 가능한 유지할 수 있도록 하기 위함이다.

수학식 1

[0053]
$$
\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} * c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} * c_{t-1} + b_f),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{sc} * X_t + W_{hc} * H_{t-1} + b_c),\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot c_t + b_o),\\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
$$

[0054] 수학식 1은 시간 디코더(223)에서 ConvLSTM이 수행하는 기능을 수학식으로 표현한 것이다. 수학식 1에서 i_t, f_t,

o_t, c_t 및 h_t 는 각각 시간 t에서 입력 게이트(input gate), 망각 게이트(forget gate), 출력 게이트(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 W_*은 다른 게이트를 연결하는 필터 행렬이고, b_*는 각 게이트에 상응하는 바이어스 벡터를 나타낸다.

[0055] 수학식 1에 나타난 바와 같이 ConvLSTM은 입력 대 상태 및 상태 대 상태 천이 시에 모두 컨볼루션 구조를 갖고 있어, 시간적 특징을 추출할 수 있을 뿐만 아니라 공간적 특징을 유지할 수 있다.

[0056] 또한 시간 디코더(223)는 순차적 디콘볼루션(deconvolution)을 통해 인가된 공간적 특징($X_{1:T}$)의 공간 해상도를 점차적으로 확대한다. 즉 시간 디코더(223)는 공간적 특징($X_{1:T}$)의 공간 구조를 유지하면서 각 프레임 간의 시간 상관에 따른 특징을 추출한다.

[0057] 이를 위해 시간 디코더(223)는 다수개의 ConvLSTM 레이어를 포함할 수 있으며, ConvLSTM 레이어 각각은 기지정된 개수의 필터를 포함할 수 있다. 도4 에서는 일예로 2개의 ConvLSTM 레이어가 각각 64개 및 32개의 필터를 포함하는 경우를 도시하였다.

[0058] 정규화기(225)는 시간 디코더(223)에서 출력되는 시공간 특징을 수학식 2에 따른 공간적 소프트 맥스(spatial softmax) 함수를 사용하여 정규화한다.

수학식 2

[0059]
$$A_{t,i} = \frac{\exp(W_i^T H_{t-1})}{\sum_j \exp(W_j^T H_{t-1})} \quad i \in 1 \cdots H \times W,$$

[0060] 수학식 2에서 H_{t-1}은 히든 상태(hidden state)를 나타내고, W_i는 위치 소프트맥스의 i번째 요소에 매핑되는 가중치이고, j는 위치를 나타낸다.

[0061] 정규화기(225)에 의해 시간 디코더(223)에서 출력되는 시공간 특징은 정규화되어 시공간 가중치($A_{1:T}$)로서 출력된다. 일예로 정규화기(225)는 시공간 가중치($A_{1:T}$)의 합이 1이되도록 정규화할 수 있다.

[0062] 특징 결합부(230)는 3D 특징($X'_{1:T}$)과 시공간 가중치($A_{1:T}$)를 수학식 3에 따라 결합하여, 감정 특징(X")을 획득한다.

수학식 3

[0063]
$$X'' = A \odot X'$$

[0064] 수학식 3에서 3D 특징($X'_{1:T}$)은 대상의 감정을 판별하기 위한 특징이고, 정규화기(225)에 의해 정규화된 시공간 가중치($A_{1:T}$)는 3D 특징($X'_{1:T}$)의 대응하는 각 영역에 대한 중요도를 지정하는 가중치로서 기능한다.

[0065] 감정값 획득부(240)는 감정 특징(X")에 대해 다시 3차원 특징을 추출하여 감정값(y)를 획득한다.

[0066] 감정값 획득부(240)는 일예로 3D 특징 추출부(210)와 유사하게 미리 학습된 3D CNN으로 구현될 수 있다. 그리고 본 실시예에서 감정값 획득부(240)는 감정값을 -1 에서 1 사이의 스칼라 값(scalar value)($y \in [-1, 1]$)으로 획득되도록 특징을 추출할 수 있으나, 이에 한정되지 않는다.

[0067] 감정값 획득부(240) 또한 효율적인 감정값을 획득하기 위해 다수개의 레이어로 구성될 수 있다. 일예로 감정값 획득부(240)는 다수개(예를 들면 4개)의 3D CNN 레이어와 다수개(예를 들면 3개)의 3D 맥스 풀링 레이어 및 다수개(예를 들면 2개)의 완전 연결 레이어(fully-connected layer)를 포함할 수 있다. 그리고 다수개의 3D CNN 레이어는 일예로 각각 32, 64, 128 및 256개의 필터를 포함할 수 있다.

[0068] 한편, 완전 연결 레이어는 단일 출력 채널을 갖고, 선형 회귀 레이어를 이용하여 감정값(y)을 획득할 수 있다.

[0069] 감정 판별부(300)는 감정값 획득부(240)에서 획득된 감정값(y)을 미리 저장된 감정값별 감정 기준에 대입함으로써, 대상의 감정을 판별한다. 상기에서 감정값(y)가 -1 에서 1 사이의 스칼라 값인 것으로 가정하였으므로, 감정값별 감정 기준은 각 감정에 대한 감정값이 -1 에서 1 사이의 연속되는 범위값으로 설정될 수 있다. 따라서, 감정 판별부(300)는 인가된 감정값(y)에 대응하는 감정을 용이하게 판별할 수 있다.

[0070] 본 실시예에서는 일예로 도2 에 도시된 2차원 감정 그래프의 각성(Arousal)과 유인가(Valence)의 2개의 축 중 유인가에 대응하는 감정값(y)을 추출한다. 그러나 이는 일예로서 경우에 따라 감정 추출부(200)는 각성에 대응하는 감정값을 추출하도록 구성될 수도 있으며, 각성 및 유인가 양쪽에 대응하는 감정값을 추출하도록 구성될 수도 있다.

[0071] 본 실시예에서 3D 특징 추출부(210)와 시공간 특징 추출부(220)의 공간 인코더(221) 및 시간 디코더(222), 그리고 감정값 획득부(240)는 각각 지정된 딥-러닝 알고리즘에 따라 미리 학습된 인공 신경망이다.

[0072] 그리고 감정 추출부(200)에 인가되는 이미지 시퀀스($I_{1:T}$)의 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)과 각 특징($X_{1:T}$, $X'_{1:T}$, X", $A_{1:T}$)는 백터 행렬(vector matrix)로 표현될 수 있다.

[0073] 결과적으로 본 실시예에 따른 감정 인식 장치는 연속되는 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)을 포함하는 이미지 시퀀스($I_{1:T}$)로부터 3차원으로 감정을 판별하기 위한 3D 특징($X'_{1:T}$)을 추출하고, 이와 동시에 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각의 시공간 주의에 기반한 특징을 추출하여 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각의 영역별 가중치($A_{1:T}$)를 획득한다. 그리고 3D 특징($X'_{1:T}$)에 영역별 가중치($A_{1:T}$)를 가중하여 감정 특징(X")을 획득하고, 감정 특징(X")으로부터 감정값(y)를 추출함으로써, 이미지 시퀀스($I_{1:T}$)에 별도의 관심 영역을 설정하지 않고서도 대상의 감정을 매우 정확하게 추출 및 판별할 수 있도록 한다.

[0074] 도5 는 도3 의 감정 인식 장치의 학습 방법을 설명하기 위한 도면이다.

[0075] 도3 및 도4 를 참조하여, 도5 의 학습 방법을 설명하면, 이미지 획득부(100)는 미리 감정값이 판별된 다수의 프레임을 포함하는 이미지 시퀀스 중 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)씩 순차적으로 감정 추출부(200)로 전달한다.

[0076] 감정 추출부(200)를 학습시키기 위해서는 다수의 이미지 시퀀스 또는 다수의 프레임이 필요하므로, 여기서는 일예로 이미지 시퀀스가 3500개의 프레임을 포함하는 경우를 도시하였다.

[0077] 그리고 이미지 획득부(100)는 이미지 시퀀스에서 순차적으로 T개씩의 프레임을 분리하여 전달하며, 이때 이미지 획득부(100)는 제1 내지 제10 프레임($I_1 \sim I_T$)을 전달하고, 이후 제2 내지 제11 프레임($I_2 \sim I_{T+1}$)을 전달하는 방식으로 전달할 수 있다.

[0078] 감정 추출부(200)의 3D 특징 추출부(210)는 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)이 포함된 이미지 시퀀스($I_{1:T}$) 전체에 대해 3D 특징($X'_{1:T}$)을 추출한다. 이와 함께 시공간 특징 추출부(220)의 공간 인코더(221)와 시간 디코더(223) 및 정규화기(225)가 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각에 대해 공간적 특징($X_{1:T}$)을 추출하고, 추출된 공간적 특징($X_{1:T}$)에 대해 다시 시공간 특징을 추출하여 정규화함으로써, 시공간 가중치($A_{1:T}$)를 획득한다.

[0079] 한편, 감정 추출부(200)의 특징 결합부(230)는 3D 특징($X'_{1:T}$)에 시공간 가중치($A_{1:T}$)를 가중하여, 감정 특징(X")을 획득하고, 감정값 획득부(240)는 획득된 감정 특징(X")에 대해 다시 3차원 측징을 추출하여, 기지정된 범위(여기서는 일예로 -1 ~ 1) 이내의 스칼라 값을 갖는 감정값(y)를 획득한다.

[0080] 여기서 3D 특징 추출부(210)와 공간 인코더(221)와 시간 디코더(223) 및 감정값 획득부(240)이 모두 학습되지 않은 인공 신경망이므로, 추출되는 3D 특징($X'_{1:T}$)과 공간적 특징($X_{1:T}$), 시공간 가중치($A_{1:T}$) 및 감정값(y)은 모두 상당한 오차를 포함한 상태이다.

[0081] 이에 획득된 감정값(y)을 해당 프레임에서 미리 판별되어 저장된 감정값과 비교하여 오차를 분석한다. 도5 의 오른쪽 그래프는 학습용으로 3500개의 프레임을 포함하는 이미지 시퀀스에서 프레임별로 획득된 감정값(y)과 미리 저장된 감정값을 나타낸다. 여기서는 감정값이 유인가 점수(Valence Score)인 경우를 나타내었으며, 청색 선은 각 프레임에 대해 획득된 감정값(y)를 나타내고, 적색 선은 미리 저장된 감정값을 나타낸다. 즉 x 축에

해당하는 특정 프레임에서 청색 선과 적색 선 사이의 차이가 오차이다.

[0082] 감정 추출부(200)는 분석된 오차가 감소하도록 3D 특징 추출부(210)와 공간 인코더(221)와 시간 디코더(223) 및 감정값 획득부(240)의 가중치 및 바이어스 벡터등을 조절하여 학습시킨다. 이때 오차는 이미지 시퀀스($I_{1:T}$)를 처리하는 순서의 역순으로 감정값 획득부(240)로부터 시간 디코더(223)와 공간 인코더(221) 및 3D 특징 추출부(210)로 전파되어, 점차로 오차를 줄이도록 학습된다.

[0083] 그리고 감정 추출부(200)는 다시 이미지 획득부(100)로부터 T개의 프레임을 인가받아, 감정값(y)을 획득하여 오차를 판별함으로써, 반복적으로 학습한다. 결과적으로 다수의 프레임을 포함하는 이미지 시퀀스에 대해 반복적으로 감정값(y)을 획득하고, 획득된 감정값(y)의 오차가 감소되도록 함으로써, 감정 인식 장치가 학습될 수 있다.

[0084] 도6 은 본 발명의 일 실시예에 따른 감정 인식 방법을 나타낸다.

[0085] 도3 및 도4 를 참조하여 도6 의 감정 인식 방법을 설명하면, 우선 이미지 획득부(100)가 T개 프레임($\{I_1, I_2, \cdots, I_T\}$)을 포함하는 이미지 시퀀스($I_{1:T}$)를 획득하여 감정 추출부(200)로 전달한다(S10).

[0086] 이에 감정 추출부(200)의 3D 특징 추출부(210)는 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)이 포함된 이미지 시퀀스($I_{1:T}$) 전체에 대해 3D 특징($X'_{1:T}$)을 추출한다(S20). 3D 특징 추출부(210)는 일예로 3D CNN으로 구현될 수 있으며, 이에 T개의 2차원 프레임($\{I_1, I_2, \cdots, I_T\}$)에서 3D 특징($X'_{1:T}$)을 추출할 수 있다.

[0087] 이와 동시에 감정 추출부(200)의 시공간 특징 추출부(220)는 T개 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각에 대해 우선 공간적 특징($X_{1:T} = \{X_1, X_2, \cdots, X_T\}$)을 추출한다(S30). 시공간 특징 추출부(220)는 일예로 2D CNN을 이용하여, T개 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각의 공간적 특징($\{X_1, X_2, \cdots, X_T\}$)을 추출할 수 있다. 이때, 시공간 특징 추출부(220)는 오버 피팅 문제를 방지하기 위해, 2D CNN으로 컨볼루션 레이어와 ReLU 레이어 및 맥스 풀링 레이어를 포함하여, 공간 해상도를 축소시킬 수 있다.

[0088] 그리고 추출된 공간적 특징($X_{1:T}$)을 가능한 유지하면서 시간적 특징을 더하기 위해, 시공간 특징을 추출한다(S40). 시공간 특징 추출부(220)는 공간적 특징($X_{1:T}$)을 유지하면서 시간적 특징을 더 추출하기 위해 일예로 ConvLSTM을 이용한다. 이때 시공간 특징 추출부(220)는 다수개의 ConvLSTM 레이어를 포함하여, 순차적 디콘볼루션함으로써, 축소된 공간 해상도를 다시 확대할 수 있다.

[0089] 그리고 감정 추출부(200)는 추출된 시공간 특징을 미리 지정된 방식으로 정규화하여, 시공간 가중치($A_{1:T}$)를 획득한다(S50).

[0090] 그리고 감정 추출부(200)의 특징 결합부(230)는 3D 특징($X'_{1:T}$)에 시공간 가중치($A_{1:T}$)를 가중하여, 감정 특징(X")을 획득한다(S60). 3D 특징($X'_{1:T}$)에 시공간 가중치($A_{1:T}$)가 가중됨으로써, T개의 프레임($\{I_1, I_2, \cdots, I_T\}$)으로부터 추출된 3D 특징($X'_{1:T}$)의 각 시공간 영역별 가중치가 상이하게 가중될 수 있다.

[0091] 이는 별도의 관심 영역이 지정되지 않더라도, 감정 추출부(200)가 T개의 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각에서 영역별 중요도를 결정할 수 있음을 의미한다.

[0092] 그리고 감정 추출부(200)의 감정값 획득부(240)는 감정 특징(X")에 대해 다시 3차원 특징을 추출하여, 기지정된 범위 이내의 스칼라 값을 갖는 감정값(y)를 획득한다(S70). 감정값 획득부(240) 또한 일예로 3D CNN으로 구현될 수 있으며, 여기서 획득된 감정값(y)는 이미지 시퀀스($I_{1:T}$)에 포함된 대상의 감정을 대표하는 값이다. 감정값 획득부(240)는 감정값을 획득하기 위해, 3D CNN 레이어와 3D 맥스 풀링 레이어 및 완전 연결 레이어를 포함할 수 있다.

[0093] 감정 판별부(300)는 미리 저장된 감정값별 대한 감정에 획득된 감정값(y)를 대입하여 비교함으로써, 대상의 감정을 판별한다(S80).

[0094] 이하에서는 본 실시예에 따른 감정 인식 장치 및 방법의 성능을 기존의 감정 인식 방법과 비교하여 설명한다.

[0095] 여기서는 본 실시예에 따른 감정 인식 장치 및 방법의 성능을 정량적으로 평가하기 위해, 평균 제곱근 오차

(Root Mean Square Error: RMSE)와 피어슨 상관 계수(Pearson Correlation Coefficient)(CC) 및 일치 상관 계수(Concordance Correlation Coefficient)(CCC)의 3가지 측정 기준을 이용하였다.

[0096] 이중 일치 상관 계수(CCC)는 수학식 4에 따라 두 변수 사이의 일치성을 측정한다.

수학식 4

[0097]
$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

[0098] 수학식 4에서 ρ는 피어슨 상관 계수이고, σ_x^2와 σ_y^2는 예측 및 측정값의 분산이며, μ_x와 μ_y는 예측 및 측정값의 평균을 나타낸다.

[0099] 그리고 성능 검증을 위한 데이터로 2015년 및 2016년 Audio/Visual Emotion recognition Challenges (이하 AV + EC)에서 채택된 RECOLA 데이터 세트와 2017년 AV + EC의 데이터 세트를 이용하였다.

[0100] 표1 은 2D CNN을 이용한 경우와 3D CNN을 이용한 경우 및 본 실시예에 따른 3D CNN과 시공간 주의(STA)를 함께 이용한 감정 인식 방법에 대한 측정 결과를 나타낸다.

표 1

[0101]

2D-CNN	3D-CNN	STA	RMSE	CC	CCC
✓			0.113	0.426	0.326
	✓		0.104	0.510	0.493
	✓	✓	0.102	0.572	0.546

[0102] 표1 의 3번째 행에 나타난 바와 같이, 본 실시예에 따른 감정 인식 방법은 3D CNN과 시공간 주의(STA)를 함께 이용함에 따라 평균 제곱근 오차(RMSE)가 줄어들었으며, 피어슨 상관 계수(CC) 및 일치 상관 계수(CCC)가 각각 0.062 및 0.053만큼 증가되었음을 알 수 있다.

[0103] 도7 은 본 실시예의 시공간 가중치를 시각화한 도면이다.

[0104] 도7 은 RECOLA 데이터 세트에 대해 시공간 특징 추출부(220)가 시공간 특징을 추출하여 획득된 시공간 가중치($A_{1:T}$)를 색상별로 구분하여 시각화한 도면이다. 도7 에서 붉은 색 영역이 가중치가 높은 영역을 나타내고, 파란색 영역은 가중치가 낮은 영역을 나타낸다.

[0105] 상기한 바와 같이, 본 발명의 실시예에서 시공간 특징 추출부(220)는 다수의 프레임($\{I_1, I_2, \cdots, I_T\}$)에 대해 공간적 특징을 추출하고, 추출된 공간적 특징을 유지하면서 시간적 특징을 더 추출함으로써, 시공간 주의 기반 특징을 추출한다. 즉 시공간 가중치($A_{1:T}$)를 획득한다. 이로 인해, 시공간 특징 추출부(220)는 도7 에 도시된 바와 같이, 별도의 관심 영역이 지정되지 않더라도, 학습된 바에 따라 다수의 프레임($\{I_1, I_2, \cdots, I_T\}$) 각각에서 감정 인식을 위한 각 영역의 가중치를 차등화시킬 수 있다.

[0106] 도7 로부터 시공간 특징 추출부(220)가 눈과 입 주위의 영역을 감정을 추정하기 위해 중요한 영역으로 스스로 판별하였음을 알 수 있다.

[0107] 도8 및 도9 는 각각 2 종류의 RECOLA 데이터 세트와 AV + EC 데이터 세트에 대해 본 실시예에 따른 감정 인식 방법을 적용하여 획득되는 감정값과 검증값을 비교한 결과를 나타낸다.

[0108] 도8 및 도9 에서 적색 선은 검증값(ground truth)를 나타내고, 청색 선은 감정값(y)를 나타낸다. 도8 및 도9 에 도시된 바와 같이, 본 발명의 실시예에 따른 감정 인식 방법에 의해 획득된 감정값(y)는 검증값과 유사하게 변동됨을 확인할 수 있다.

[0109] 그리고 표2 및 표3 에서는 각각 RECOLA 데이터 세트와 AV + EC 데이터 세트에 대한 본 실시예에 따른 감정 인식

결과를 다른 감정 인식 방법과 비교하였다.

표 2

[0110]

Method	RMSE	CC	CCC
Baseline [26]	0.117	0.358	0.273
CNN [1]	0.113	0.426	0.326
CNN + RNN (≈ 1 sec.) [1]	0.111	0.501	0.474
CNN + RNN (≈ 4 sec.) [1]	0.108	0.544	0.506
LGBP-TOP + LSTM [29]	0.114	0.430	0.354
LGBP-TOP + Bi-Dir. LSTM [15]	0.105	0.501	0.346
LGBP-TOP + LSTM + ϵ-loss [30]	0.121	0.488	0.463
CNN + LSTM + ϵ-loss [30]	0.116	0.561	0.538
3D-CNN + STA (≈ 4 sec.)	**0.102**	**0.572**	**0.546**

표 3

[0111]

Method	RMSE	CC	CCC
Baseline [31]	-	-	0.400
CNN [1]	0.114	0.564	0.528
CNN + RNN (≈ 4 sec.) [1]	0.104	0.616	0.588
3D-CNN + STA (≈ 4 sec.)	**0.099**	**0.638**	**0.612**

[0112] 상기 표2 및 표3 에 나타난 바와 같이, 본 실시예에 따른 감정 인식 방법은 기존의 다른 감정 인식 방법에 비해, 가장 낮은 평균 제곱근 오차(RMSE)를 나타내는 반면, 피어슨 상관 계수(CC) 및 일치 상관 계수(CCC)는 가장 높게 나타남을 확인할 수 있다. 즉 감정 인식 성능이 매우 우수함을 확인할 수 있다.

[0113] 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.

[0114] 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.

[0115] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0116] 100: 이미지 획득부 200: 감정 추출부

300: 감정 판별부 210: 3D 특징 추출부

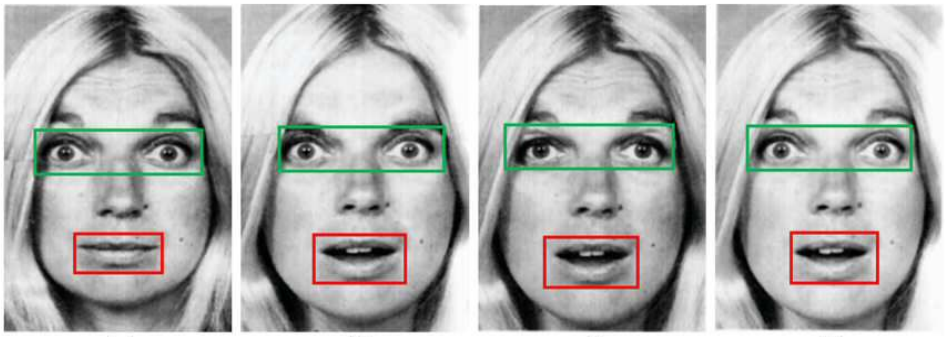
220: 시공간 특징 추출부 230: 특징 결합부

240: 감정값 획득부 221: 공간 인코더

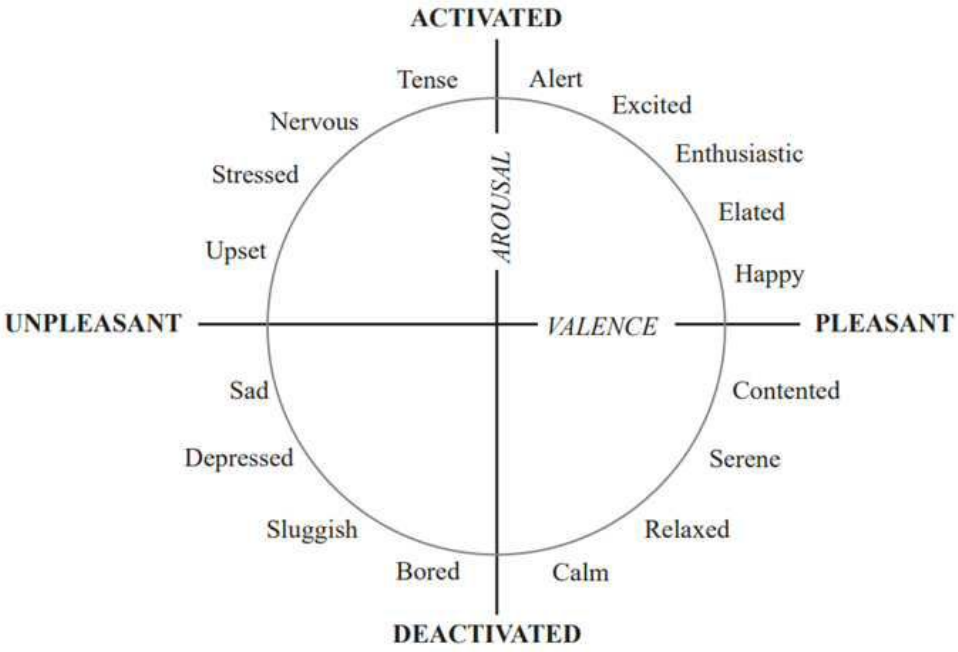
223: 시간 디코더 225: 정규화부

도면

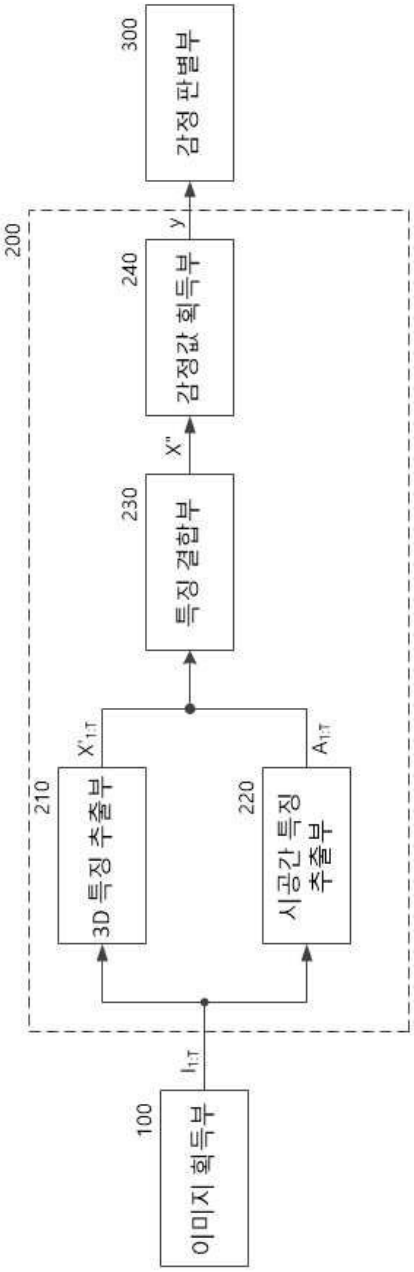
도면1


(a)
(b)
(c)
(d)

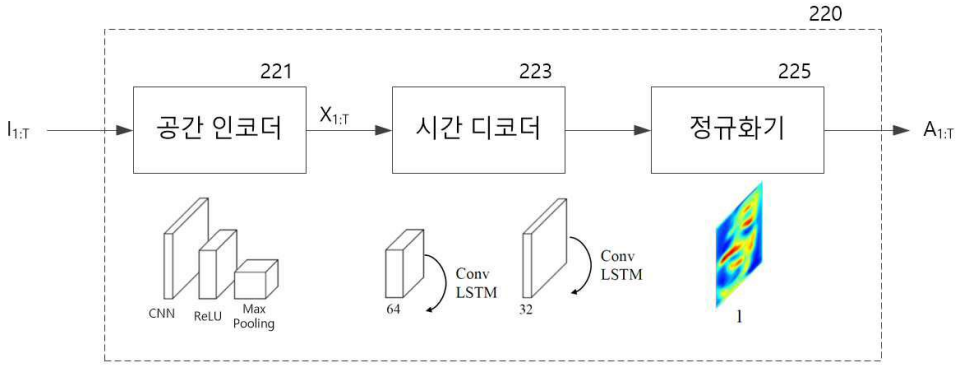
도면2


ACTIVATED
Tense
Alert
Excited
Nervous
Enthusiastic
Stressed
AROUSAL
Elated
Upset
Happy
UNPLEASANT
VALENCE
PLEASANT
Sad
Contented
Depressed
Serene
Sluggish
Relaxed
Bored
Calm
DEACTIVATED

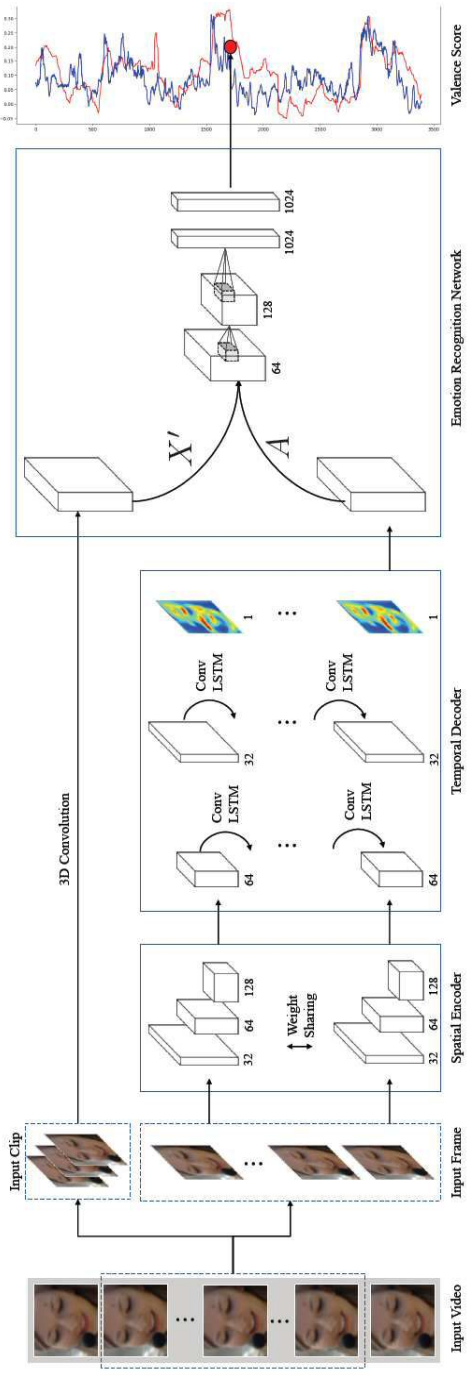
도면3


300
감정 판별부
y
200
240
감정값 획득부
X''
230
특징 결합부
X1:T
A1:T
210
3D 특징 추출부
220
시공간 특징 추출부
I1:T
100
이미지 획득부

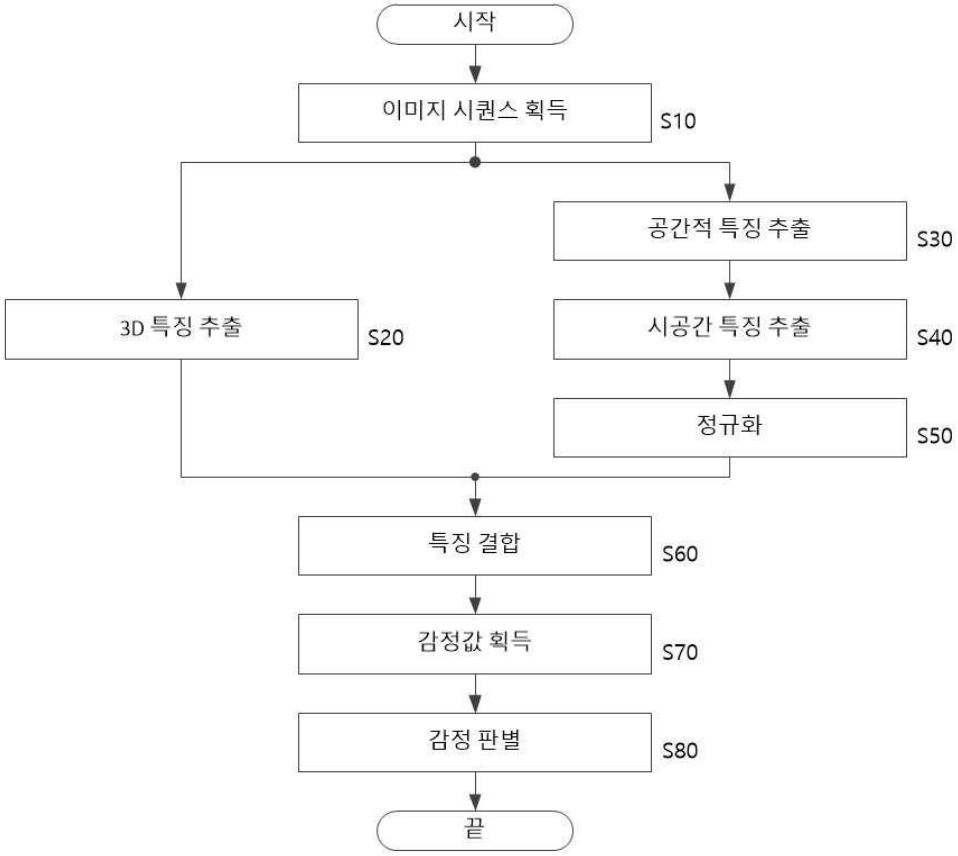
도면4


220
I1:T
221
공간 인코더
X1:T
223
시간 디코더
225
정규화기
A1:T
CNN
ReLU
Max Pooling
64
Conv LSTM
32
Conv LSTM
1

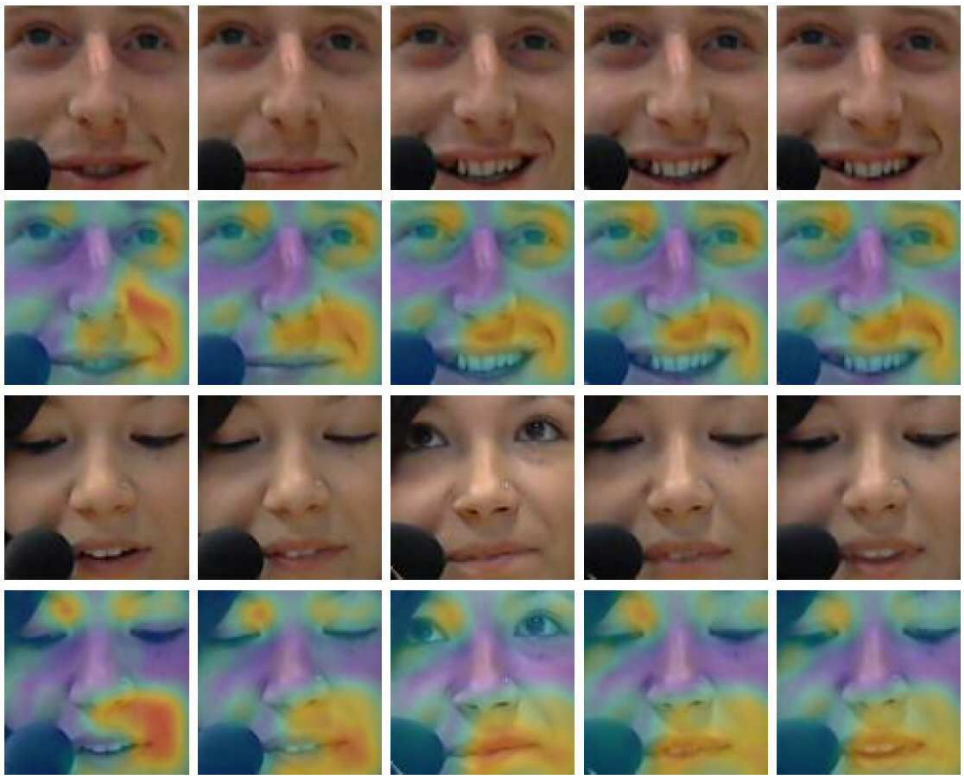
도면5


Valence Score
Emotion Recognition Network
1024
1024
128
64
X'
A
Temporal Decoder
Conv LSTM
Conv LSTM
Conv LSTM
Conv LSTM
32
32
64
64
1
1
3D Convolution
Spatial Encoder
Weight Sharing
32
64
128
32
64
128
Input Clip
Input Frame
Input Video

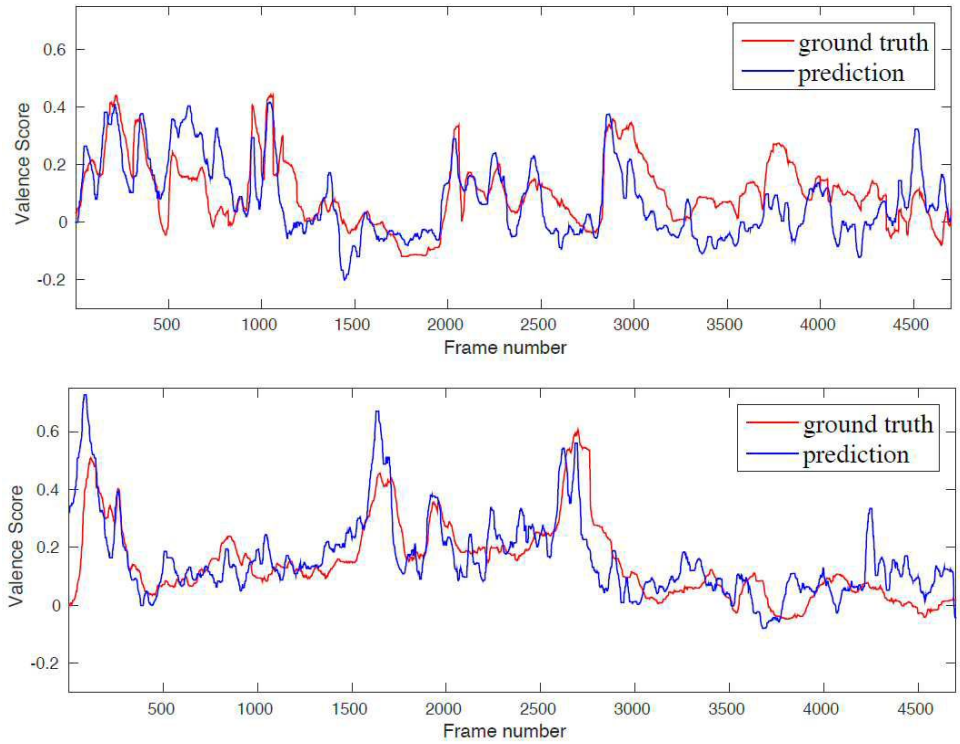
도면6


시작
이미지 시퀀스 획득
S10
공간적 특징 추출
S30
3D 특징 추출
S20
시공간 특징 추출
S40
정규화
S50
특징 결합
S60
감정값 획득
S70
감정 판별
S80
끝

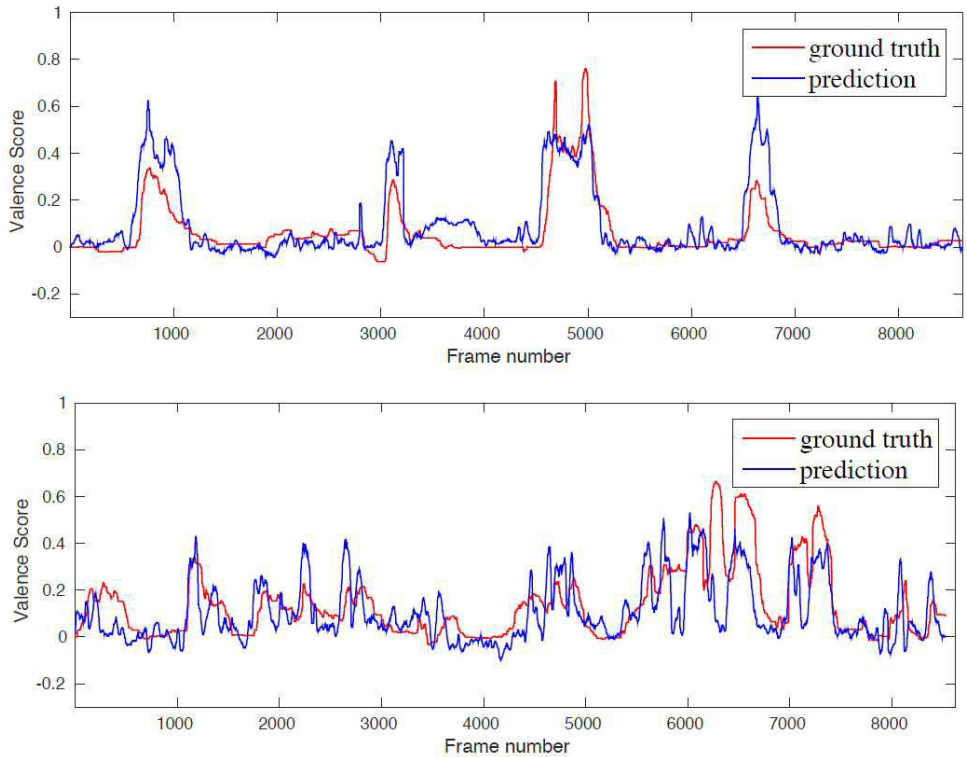
도면7



도면8


ground truth
prediction
Valence Score
Frame number
0.6
0.4
0.2
0
-0.2
500
1000
1500
2000
2500
3000
3500
4000
4500

도면9


ground truth
prediction
Valence Score
Frame number
1
0.8
0.6
0.4
0.2
0
-0.2
1000
2000
3000
4000
5000
6000
7000
8000