(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0152071
(43) 공개일자 2024년10월21일

(51) 국제특허분류(Int. Cl.)
G06V 40/16 (2022.01) *G06N 20/20* (2019.01)
G06V 10/25 (2022.01) *G06V 10/26* (2022.01)
G06V 10/30 (2022.01) *G06V 10/774* (2022.01)
G06V 10/80 (2022.01) *G06V 20/40* (2022.01)
(52) CPC특허분류
G06V 40/174 (2022.01)
G06N 20/20 (2021.08)
(21) 출원번호 10-2023-0048172
(22) 출원일자 2023년04월12일
심사청구일자 2023년04월12일

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 갯벌로 27, 인천대학교 이노베이션센터(송도동)
(72) 발명자
전광길
인천광역시 연수구 해송로 70, 209동 805호(송도동, 송도웰카운티2단지)
(74) 대리인
특허법인충정

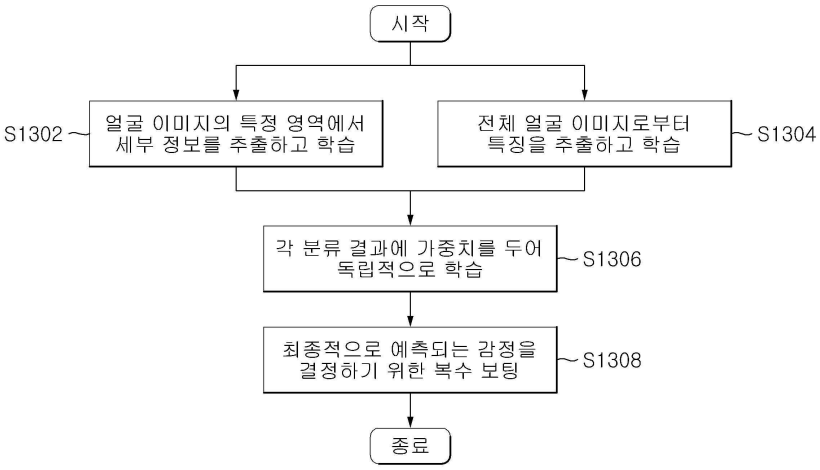
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **안면 인식을 위한 네트워크 조합 방법 및 장치**

(57) 요 약

본 발명은 안면 인식을 위한 네트워크 조합 방법 및 장치에 관한 것으로, 다양한 표정의 얼굴 이미지 데이터셋을 입력 받아 전처리한 이미지를 감정과 관련된 특정 영역에서의 세부 정보를 추출하고 학습하는 로컬 인식 단계; 상기 이미지의 얼굴 전체 정보를 추출하고 학습하는 공유 표현 글로벌 앙상블 단계; 각 분류기의 가중치를 할당하는 랭크 어텐션 단계; 및 예측되는 감정을 최종적으로 결정하기 위한 복수 보팅 단계;를 포함하는 것을 특징으로 한다.

대 표 도 - 도13



(52) CPC특허분류
G06V 10/25 (2023.08)
G06V 10/26 (2023.08)
G06V 10/30 (2023.08)
G06V 10/774 (2023.08)
G06V 10/809 (2023.08)
G06V 10/817 (2023.08)
G06V 20/49 (2022.01)
G06V 40/169 (2022.01)
G06V 40/171 (2022.01)

명 세 서

청구범위

청구항 1

다양한 표정의 얼굴 이미지 데이터셋을 입력 받아 전처리한 이미지를 감정과 관련된 특정 영역에서의 세부 정보를 추출하고 학습하는 로컬 인식 단계;

상기 이미지의 얼굴 전체 정보를 추출하고 학습하는 공유 표현 글로벌 앙상블 단계;

각 분류기의 가중치를 할당하는 랭크 어텐션 단계; 및

예측되는 감정을 최종적으로 결정하기 위한 복수 보팅 단계;를 포함하는 것을 특징으로 하는, 안면 인식을 위한 네트워크 조합 방법.

청구항 2

제1항에 있어서,

상기 로컬 인식 단계는, 두 정점의 좌표를 기반으로 왼쪽 눈 ROI(Region of interest), 오른쪽 눈 ROI, 코와 입 ROI 세 개의 영역으로 분리함으로써 얼굴 이미지를 구분하는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 방법.

청구항 3

제1항에 있어서,

상기 로컬 인식 단계는 약분류 단계와 강분류 단계를 포함하는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 방법.

청구항 4

제1항에 있어서,

상기 공유 표현 글로벌 앙상블 단계에서 분기(branch)를 시작하는 레벨은 분기 레벨 2 내지 분기 레벨 4 중 하나를 포함하는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 방법.

청구항 5

제1항에 있어서,

상기 전처리는 비디오 스트리밍 데이터에서 처음 프레임 및 마지막 3개의 프레임을 선택하여 회색조(grayscale) 사진으로 변환하되, 상기 마지막 3개는 GEMSR-4(공유 레이어와 4개의 앙상블 분기만 포함)의 입력으로 사용될 감정 레이블 태그가 붙고, Viola & Jones의 알고리즘으로 얼굴을 분할함으로써 얼굴을 감지하고 정렬하여 배경 또는 얼굴이 아닌 영역을 제거할 수 있는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 방법.

청구항 6

제1항에 있어서, 상기 랭크 어텐션 단계에서는 가중치 벡터 S는

$$S = F_{rank}(Z, W) = \sigma\left(W_2 \delta\left(W_1 Z\right)\right)$$

에 기반하여 계산될 수 있고,

여기서 $W_1, W_2 \in \mathbb{R}^{C \times C}$, δ는 ReLU, σ는 Sigmoid 함수,

$Z^m = [z_1^m, z_2^m, \ldots z_i^m]$는 m번째 분기의 초기 분류 값을 의미하는, 안면 인식을 위한 네트워크 조합 방법.

청구항 7

제6항에 있어서, 상기 복수 보팅 단계에서는

초기 분류 결과 값은 $Z^m = \mathrm{F}_{CLS}(Z, W) = WZ$

에 기반하여 계산될 수 있고,

여기서 매개 변수들은 $\tilde{z}_i^m = F_{\text{scale}}(z_i^m, s_i) = s_i \cdot z_i^m$ 와 같이 가중치 s_i를 부과할 수 있으며,

$$Z_{\text{final}} = c_{\arg_i \max} \sum_{m=1}^{M} \tilde{Z}^m$$

에 기반하여 복수 보팅의 결과 값(Z_{final})을 계산할 수 있는데, c_i는 주어진 감정 라벨을 의미하고, argmax(arguments of maximum)는 목적 함수(Z)를 최대화시키는 입력 값을 찾는 역할을 수행하는, 안면 인식을 위한 네트워크 조합 방법.

청구항 8

다양한 표정의 얼굴 이미지 데이터셋을 입력 받아 전처리한 이미지를 감정과 관련된 특정 영역에서의 세부 정보를 추출하고 학습하는 로컬 인식 모듈(Local Perception Module, LPM); 상기 이미지의 얼굴 전체 정보를 추출하고 학습하는 공유 표현 글로벌 앙상블 모듈(Global Ensemble Module with Shared Representation, GEMSR); 각 분류기의 가중치를 할당하는 랭크 어텐션 유닛 모듈(Rank Attention Unit, RAU) 및 예측되는 감정을 최종적으로 결정하기 위한 복수 보팅(plurality voting) 모듈;을 포함하는 것을 특징으로 하는, 안면 인식을 위한 네트워크 조합 장치.

청구항 9

제8항에 있어서,

상기 로컬 인식 모듈은, 두 정점의 좌표를 기반으로 왼쪽 눈 ROI(Region of interest), 오른쪽 눈 ROI, 코와 입 ROI 세 개의 영역으로 분리함으로써 얼굴 이미지를 구분하는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 장치.

청구항 10

제8항에 있어서,

상기 로컬 인식 모듈은 약분류기(weak classifier)와 강분류기(strong classifier)를 포함하는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 장치.

청구항 11

제8항에 있어서,

상기 공유 표현 글로벌 앙상블 모듈에서 분기(branch)를 시작하는 레벨은 분기 레벨 2 내지 분기 레벨 4 중 하나를 포함하는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 장치.

청구항 12

제8항에 있어서,

상기 전처리는 비디오 스트리밍 데이터에서 처음 프레임 및 마지막 3개의 프레임을 선택하여 회색조(grayscale)

사진으로 변환하되, 상기 마지막 3개는 GEMSR-4(공유 레이어와 4개의 앙상블 분기만 포함)의 입력으로 사용될 감정 레이블 태그가 붙고, Viola & Jones의 알고리즘으로 얼굴을 분할함으로써 얼굴을 감지하고 정렬하여 배경 또는 얼굴이 아닌 영역을 제거할 수 있는 것을 특징으로 하는 안면 인식을 위한 네트워크 조합 장치.

청구항 13

제8항에 있어서,

상기 랭크 어텐션 모듈에서는

가중치 벡터 S를 $S = F_{rank}(Z, W) = \sigma\left(W_2\delta\left(W_1 Z\right)\right)$ 에 기반하여 표현될 수 있고,

여기서 $W_1, W_2 \in \mathbb{R}^{C\times C}, \delta, \sigma$ 는 RELU, Sigmoid 함수,

$Z^m = [z_1^m, z_2^m, \ldots z_i^m]$ 는 m번째 분기의 초기 분류 값을 의미하는, 안면 인식을 위한 네트워크 조합 장치.

청구항 14

제13항에 있어서, 상기 복수 보팅 모듈에서는

초기 분류 결과 값은 $Z^m = \mathrm{F}_{CLS}(Z, W) = WZ$

에 기반하여 계산될 수 있고,

여기서 매개 변수들은 $\tilde{z}_i^m = F_{\text{scale}}\left(z_i^m, s_i\right) = s_i \cdot z_i^m$ 와 같이 가중치 s_i를 부과할 수 있으며,

$$Z_{\text{final}} = c_{\arg_i \max} \sum_{m=1}^{M} \tilde{Z}^m$$

에 기반하여 복수 보팅의 결과 값(Z_{final})을 계산할 수 있는데, c_i는 주어진 감정 라벨을 의미하고, argmax(arguments of maximum)는 목적 함수(Z)를 최대화시키는 입력 값을 찾는 역할을 수행하는, 안면 인식을 위한 네트워크 조합 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 얼굴 표정 인식(Facial Expression Recognition, FER)을 위해 인공지능을 활용하여 개발한 글로벌 및 로컬 퓨전 앙상블 네트워크(Global Local Fusion Ensemble Network, GLFEN) 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 인간의 의사소통에서 주고받는 정보 중 표정은 인간이 감정을 표현하는 가장 직관적이고 효과적인 방법 중 하나로써, 전달된 정보의 효과 중 얼굴 표정이 55%를 차지하며 언어와 음성이 각 7%와 38%를 차지한다.

[0003] 얼굴 표정을 기반으로 한 자동 감정 인식은 수익성이 있고 광범위한 응용 가능성을 갖고 있는데, 얼굴 표정 인식(Facial Expression Recognition, FER)은 로봇 공학, 교육, 정서적 건강과 같은 인간-컴퓨터 상호 작용(Human-Computer Interaction, HCI) 응용 분야에서 큰 주목을 받고 있다. 따라서 컴퓨터가 인간의 감정 상태를 효율적이고 효과적으로 추론하도록 하는 방법은 자연스러운 인간-컴퓨터 상호 작용을 달성하는 데 필수적인 방법이다.

[0004] 일반적으로 자동 얼굴 표정 인식 시스템의 핵심은 특징 추출 기능으로, 원본 얼굴 이미지에서 대표적인 특징을

추출한 후 분류하는 것을 목표로 한다. 예를 들면, 앙상블 학습 전략과 CNN을 활용하여 분류 결과를 얻을 수 있는데, 앙상블 학습 전략은 분류 작업에 널리 적용되는 것으로, 표본으로 추정한 회귀식과 실제 관측값의 차이를 일컫는 잔차(residual)의 일반화 오류를 줄일 수 있을 뿐만 아니라 과적합(overfitting) 문제도 해결 가능하다는 특징을 갖는다. CNN(Convolutional Neural Networks)은 강력한 특징 표현 능력으로 인해 다양한 컴퓨터 비전 작업에 널리 사용되고 있다. 한편, 위 방법은 깊은 앙상블 네트워크를 학습하는 데 비용이 많이 들고 low-level features의 중복(Redundancy)이 발생한다는 문제점이 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) KR 10-2021-0087749 A

발명의 내용

해결하려는 과제

[0006] 이와 같은 문제점을 해결하기 위하여, 본 발명은 효율적인 Global and Local Fusion Ensemble Network(GLFEN)을 제공하는데 그 목적이 있다.

[0007] 얼굴 이미지 배치가 주어지면 GEMSR(Global Ensemble Module with Shared Representations)은 먼저 전체 얼굴 정보를 추출하는 데 사용된다.

[0008] 그리고 LPM(Local Perception Module)은 감정과의 관련성이 높은 얼굴 영역(예: 눈, 입 주변 영역)의 반응을 캡처하는 데 적용되어 특징 추출의 세부 정보를 제공한다.

[0009] RAU(Rank Attention Unit)는 각 분류기의 가중치를 학습할 수 있으며, 감정을 예측하기 위한 최종단계에서 복수 보팅(voting)을 수행함으로써 학습된 데이터를 융합(fuse)할 수 있다.

[0010] 본 발명에서 제시하는 GLFEN 모델은 상기 GEMSR, LPM, RAU 및 복수 보팅으로 구성됨으로써 중복(Redundancy)을 줄일 수 있고, 앙상블 모듈의 속도를 보장할 수 있으며, 결과적으로 네트워크의 견고성을 향상시킬 수 있다.

과제의 해결 수단

[0011] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법은, 다양한 표정의 얼굴 이미지 데이터셋을 입력 받아 전처리한 이미지를 감정과 관련된 특정 영역에서의 세부 정보를 추출하고 학습하는 로컬 인식 단계; 상기 이미지의 얼굴 전체 정보를 추출하고 학습하는 공유 표현 글로벌 앙상블 단계; 각 분류기의 가중치를 할당하는 랭크 어텐션 단계; 및 예측되는 감정을 최종적으로 결정하기 위한 복수 보팅 단계;를 포함할 수 있다.

[0012] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 중 상기 로컬 인식 단계는, 두 정점의 좌표를 기반으로 왼쪽 눈 ROI(Region of interest), 오른쪽 눈 ROI, 코와 입 ROI 세 개의 영역으로 분리함으로써 얼굴 이미지를 구분하는 것을 포함할 수 있다.

[0013] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 중 상기 로컬 인식 단계는 약분류 단계와 강분류 단계를 포함할 수 있다.

[0014] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 중 상기 공유 표현 글로벌 앙상블 단계에서 분기(branch)를 시작하는 레벨은 분기 레벨 2 내지 분기 레벨 4 중 하나를 포함할 수 있다.

[0015] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 중 상기 전처리는 비디오 스트리밍 데이터에서 처음 프레임 및 마지막 3개의 프레임을 선택하여 회색조(grayscale) 사진으로 변환하되, 상기 마지막 3개는 GEMSR-4(공유 레이어와 4개의 앙상블 분기만 포함)의 입력으로 사용될 감정 레이블 태그가 붙고, Viola & Jones의 알고리즘으로 얼굴을 분할함으로써 얼굴을 감지하고 정렬하여 배경 또는 얼굴이 아닌 영역을 제거할 수 있는 것을 포함할 수 있다.

[0016] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 중 상기 랭크 어텐션 단계에서는 가중치 벡

터 S는

[0017] $$S = F_{rank}(Z, W) = \sigma\left(W_2 \delta\left(W_1 Z\right)\right)$$

[0018] 에 기반하여 계산될 수 있고,

[0019] 여기서 $W_1, W_2 \in \mathbb{R}^{C \times C}$, δ는 ReLU, σ는 Sigmoid 함수,

[0020] $Z^m = \left[z_1^m, z_2^m, \ldots z_i^m\right]$는 m번째 분기의 초기 분류 값을 의미하는 것을 포함할 수 있다.

[0021] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 중 상기 복수 보팅 단계에서는

[0022] 초기 분류 결과 값은 $Z^m = \mathrm{F}_{CLS}(Z, W) = WZ$

[0023] 에 기반하여 계산될 수 있고,

[0024] 여기서 매개 변수들은 $\tilde{z}_i^m = F_{\text{scale}}\left(z_i^m, s_i\right) = s_i \cdot z_i^m$ 와 같이 가중치 s_i를 부과할 수 있으며,

[0025] $$Z_{\text{final}} = c_{\arg_i \max} \sum_{m=1}^{M} \tilde{Z}^m$$

[0026] 에 기반하여 복수 보팅의 결과 값(Z_{final})을 계산할 수 있는데, c_i는 주어진 감정 라벨을 의미하고, argmax(arguments of maximum)는 목적 함수(Z)를 최대화시키는 입력 값을 찾는 역할을 수행하는 것을 포함할 수 있다.

[0027] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치는, 다양한 표정의 얼굴 이미지 데이터셋을 입력 받아 전처리한 이미지를 감정과 관련된 특정 영역에서의 세부 정보를 추출하고 학습하는 로컬 인식 모듈(Local Perception Module, LPM); 상기 이미지의 얼굴 전체 정보를 추출하고 학습하는 공유 표현 글로벌 앙상블 모듈(Global Ensemble Module with Shared Representation, GEMSR); 각 분류기의 가중치를 할당하는 랭크 어텐션 유닛 모듈(Rank Attention Unit, RAU) 및 예측되는 감정을 최종적으로 결정하기 위한 복수 보팅(plurality voting) 모듈;을 포함하는 것을 특징으로 하는 것을 포함할 수 있다.

[0028] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치 중 상기 로컬 인식 모듈은, 두 정점의 좌표를 기반으로 왼쪽 눈 ROI(Region of interest), 오른쪽 눈 ROI, 코와 입 ROI 세 개의 영역으로 분리함으로써 얼굴 이미지를 구분하는 것을 포함할 수 있다.

[0029] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치 중 상기 로컬 인식 모듈은 약분류기(weak classifier)와 강분류기(strong classifier)를 포함할 수 있다.

[0030] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치 중 상기 공유 표현 글로벌 앙상블 모듈에서 분기(branch)를 시작하는 레벨은 분기 레벨 2 내지 분기 레벨 4 중 하나를 포함할 수 있다.

[0031] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치 중 상기 전처리는 비디오 스트리밍 데이터에서 처음 프레임 및 마지막 3개의 프레임을 선택하여 회색조(grayscale) 사진으로 변환하되, 상기 마지막 3개는 GEMSR-4(공유 레이어와 4개의 앙상블 분기만 포함)의 입력으로 사용될 감정 레이블 태그가 붙고, Viola & Jones의 알고리즘으로 얼굴을 분할함으로써 얼굴을 감지하고 정렬하여 배경 또는 얼굴이 아닌 영역을 제거할 수 있는 것을 포함할 수 있다.

[0032] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치 중 상기 랭크 어텐션 모듈에서는

[0033] 가중치 벡터 S는

[0034] $$S = F_{rank}(Z, W) = \sigma\left(W_2 \delta\left(W_1 Z\right)\right)$$

[0035] 에 기반하여 계산될 수 있고,

[0036] 여기서 $W_1, W_2 \in \mathbb{R}^{C \times C}$, δ는 ReLU, σ는 Sigmoid 함수,

[0037] $Z^m = [z_1^m, z_2^m, \ldots z_i^m]$는 m번째 분기의 초기 분류 값을 의미하는 것을 포함할 수 있다.

[0038] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치 중 상기 복수 보팅 모듈에서는

[0039] 초기 분류 결과 값은 $Z^m = F_{CLS}(Z, W) = WZ$

[0040] 에 기반하여 계산될 수 있고,

[0041] 여기서 매개 변수들은 $\tilde{z}_i^m = F_{\text{scale}}(z_i^m, s_i) = s_i \cdot z_i^m$ 와 같이 가중치 s_i를 부과할 수 있으며,

$$Z_{\text{final}} = c_{\arg_i \max} \sum_{m=1}^{M} \tilde{Z}^m$$

[0042]

[0043] 에 기반하여 복수 보팅의 결과 값(Z_{final})을 계산할 수 있는데, c_i는 주어진 감정 라벨을 의미하고, argmax(arguments of maximum)는 목적 함수(Z)를 최대화시키는 입력 값을 찾는 역할을 수행하는 것을 포함할 수 있다.

발명의 효과

[0044] 본 발명은 실시간으로 인간과 컴퓨터가 상호작용(HCI)하는데 유용한 얼굴 표정 인식을 위한 글로벌 및 로컬 퓨전 앙상블 네트워크를 제시한다. 프레임워크는 로컬 인식 모듈, 글로벌 앙상블 모듈, 어텐션 모듈 및 복수 보팅 모듈을 포함한 부분으로 구성된다.

[0045] 전체 얼굴 이미지로부터 특징을 추출하고 학습하여 전역 정보를 학습하고, 서로 다른 데이터 사이에 존재하는 공통의 정보인 Shared Representation이 있는 앙상블 모듈을 적용하여 고유한 특징(features)을 학습함으로써, 네트워크 매개 변수 수와 앙상블의 중복성을 줄일 수 있다.

[0046] 얼굴 이미지의 특정 영역에서 세부 정보를 추출하고 학습하여 추출된 정보의 무결성을 보장하고, 얼굴 간의 상관성에 초점을 맞출 수 있다.

[0047] 각 분류 결과에 가중치를 두고 독립적으로 학습하고, 최종적으로 복수 보팅(voting)을 함으로써 서로 다른 개별 학습자에 대한 분류 결과의 중요성 또는 신뢰성을 보장할 수 있다.

[0048] 결과적으로 본 발명은 앙상블의 견고성을 향상시킬 뿐만 아니라 공개 데이터 세트에 대한 정확도 측면에서 경쟁력 있는 성과를 달성한다는 것을 확인할 수 있다.

도면의 간단한 설명

[0049] 도 1은 본 발명에서 제안하는 GLFEN(Global and Local Fusion Ensemble Network) 및 그 하위 모듈의 프레임워크에 관한 것이다.

도 2는 왼쪽에는 GEMSR-4 Lv1.3 아키텍처와 오른쪽에는 다양한 분기 수준 그림을 나타낸 것이다.

도 3은 얼굴의 영역 분해로, ROI의 각 점은 특정 좌표에 해당한다.

도 4는 왼쪽은 원래 분류(CLS) 네트워크 스키마와 오른쪽은 RAU를 나타낸다.

도 5는 CK+(Extended Cohn-Kanade) 데이터셋 및 FER+의 이미지 샘플(8개 클래스)을 나타낸다.

도 6은 JAFFE 및 RAF-DB의 이미지 샘플(7개 클래스)을 나타낸다.

도 7은 오리지널 네트워크의 branch level이 높아지는 동안의 CK+ 데이터 세트의 정확도를 나타낸다.

도 8은 10개의 폴드를 만들어 교차 검증(10-fold Cross Validation)하는 네트워크와 기준선(baseline) 사이의 각 폴드(fold)에 대한 평균 정확도(%) 비교한 것으로 파란색 막대는 ESR-4를 나타내고 주황색 막대는 GLFEN을 나타낸다.

도 9는 FER+에 대한 앙상블 예측의 정규화된 컨퓨전 매트릭스(confusion matrix)를 나타낸다.

도 10은 다양한 얼굴 이미지를 가진 대규모 데이터베이스인 RAF-DB (Real-world Affective Faces Database)의 일부 가려진 샘플을 나타낸다.

도 11은 RAF-DB에 대한 앙상블 예측의 정규화된 컨퓨전 매트릭스(confusion matrix)를 나타낸다.

도 12는 Grad-CAM의 시각화를 나타낸 것으로 AM은 어텐션(attention) 메커니즘(즉, RAU)을 나타내고, 유색 영역은 출력 활성화에 가장 크게 기여하는 얼굴 특징을 나타낸다.

도 13은 본 발명의 일 실시 예에 따라 안면 인식을 위한 네트워크 조합을 개발하기 위한 일반적인 절차를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0050] 본 발명의 목적, 특정한 장점들 및 신규한 특징들은 첨부된 도면들과 연관되어지는 이하의 상세한 설명과 바람직한 실시예들로부터 더욱 명백해질 것이다. 이때, 각각의 도면에서 동일한 구성요소는 가능한 동일한 부호로 나타낸다. 또한, 이미 공지된 기능 및/또는 구성에 대한 상세한 설명은 생략한다.

[0051] 이하에 개시된 내용은, 다양한 실시 예에 따른 동작을 이해하는데 필요한 부분을 중점적으로 설명하며, 그 설명의 요지를 흐릴 수 있는 요소들에 대한 설명은 생략한다. 또한 도면의 일부 구성요소는 과장되거나 생략되거나 또는 개략적으로 도시될 수 있다. 각 구성요소의 크기는 실제 크기를 전적으로 반영하는 것이 아니며, 따라서 각각의 도면에 그려진 구성요소들의 상대적인 크기나 간격에 의해 여기에 기재되는 내용들이 제한되는 것은 아니다.

[0052] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시 예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

[0053] 또한, 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되는 것은 아니며, 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0055] 이하에서는, 첨부된 도면들을 참조하여, 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법 및 장치에 대해 설명하기로 한다.

[0056] 본 발명의 GLFEN(Global and Local Fusion Ensemble Network) 흐름도는 도 13에 나와 있다.

[0057] 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 방법에는 4가지 주요 단계가 있다. 특징 추출 단계(단계 S1302, 단계S1304); 가중치 어텐션 단계(단계 S1306); 및 복수 보팅 단계(단계 S1308)를 포함하는 것을 특징으로 한다.

[0058] 도 13을 참조하면, 단계 S1302에서, 로컬 인식 모듈(Local Perception Module, LPM)(102)은 얼굴 이미지의 특정 영역에서 세부 정보를 추출하고 학습한다.

[0059] 단계 S1304에서, 공유 레이어 글로벌 앙상블 모듈(Global Ensemble Module with Shared Representation, GEMSR)(104)은 전체 얼굴 이미지로부터 특징을 추출하고 학습한다.

[0060] 단계 S1306에서, 가중치 어텐션(Rank Atttetnion Unit, RAU) 모듈(106)은 각 분류 결과에 가중치를 두어 독립적으로 학습한다.

[0061] 단계 S1308에서, 복수의 보팅(Plurality Voting) 모듈(108)은 최종적으로 예측되는 감정을 결정하기 위한 복수 보팅을 한다.

[0062] 한편, 도 1은 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치의 블록도이다.

[0063] 도 1에 도시된 본 발명의 일 실시예에 의한 안면 인식을 위한 네트워크 조합 장치는, 다양한 표정을 짓고있는 얼굴 이미지 데이터셋을 입력 받아 전처리한 이미지에서 감정과 관련된 특정 영역에서 세부정보를 추출하고 분류하는 로컬 인식 모듈(Local Perception Module, 이하 'LPM'라 칭함.)(102), 상기 이미지의 얼굴 전체 정보를 추출하는 공유 글로벌 앙상블 모듈(Global Ensemble Module with Shared Representation, 이하 'GEMSR'라 칭함.)(104), 각 분류기의 가중치를 할당하는 랭크 어텐션 유닛 모듈(Rank Attention Unit, 이하 'RAU'라 칭함.)(106), 예측되는 감정을 최종적으로 결정하기 위한 복수 보팅(Voting) 모듈(108)을 포함한다.

[0064] 도 1은 제안된 GLFEN(Global and Local Fusion Ensemble Network) 및 그 하위 모듈의 프레임워크를 나타내는 도면이다.

[0065] 도 1을 참조하면, 각 모듈은 인풋 이미지에서 특징(features)을 추출하여 학습한 후에 어느 하나의 감정으로 분류할 수 있다. 그 과정은 다음과 같이 세 가지 전략으로 구성될 수 있다.

[0066] 첫 번째로 두 개의 빌딩 블록(building blocks)을 포함하는 GEMSR(104)을 사용하는 것일 수 있다.

[0067] 도 2를 참조하면, 오른쪽에는 레벨 1, 레벨 3, 레벨 5의 다양한 분기 수준(branch level)을 나타낸 것이고, 왼쪽에는 그 중 GEMSR-4 분기 레벨 3 아키텍처를 나타낸 것이다.

[0068] 본 발명의 일 실시예로, 도 2에서의 회색 블록과 같이 low level과 middle level의 학습을 위한 일련의 컨볼루션 레이어(convolution layer)로 구성될 수 있다. 뒷 부분의 파란색 분기(branch)는 정보들의 특징(features)을 공유하는 글로벌 앙상블 모듈을 구성할 수 있다.

[0069] 이러한 효율적인 네트워크 구축의 이점은 각 분기(branch)가 고유한 특징(feature)을 학습할 수 있도록 할 뿐만 아니라 공유 계층(shared layer)을 통해 low 및 middle level 정보의 중복(redundancy)을 감소시킬 수 있다.

[0070] 두 번째로 LPM(Local Perception Module)(102)을 로컬 분류기로 사용할 수 있다. 이 모듈은 인풋 얼굴 이미지의 여러 주요 영역에서 로컬 세부 정보를 추출하고 학습할 수 있다. 따라서 이러한 모듈에 의해 글로벌 정보가 보완됨으로써 앙상블의 다양성을 증가시킬 수 있다.

[0071] 세 번째로 최종적으로 감정을 판단하기 이전에, 분류 결과에 대한 적절한 퓨전을 달성하기 위해 어텐션 메커니즘(attention mechanism)을 활용할 수 있다. 상기 두 분기 단계에서 완전한 정보를 학습할 수 있더라도 최종 분류 결과에 미치는 영향은 여전히 서로 다를 수 있다. 따라서 본 발명은 전체 네트워크의 연산량에 부담이 거의 없고, 각 분류기의 가중치를 독립적으로 학습하는 랭크 어텐션 유닛(rank attention units,RAU)(106)을 설계할 수 있다.

[0072] 실제로 GEMSR(104)에서 분기를 시작하는 레벨은 중복성(redundancy)과 앙상블 다양성에 상당한 영향을 미치기 때문에 신중하게 고려할 가치가 있는데, 도 2의 레벨 1과 같이 분기를 너무 일찍 시작하면 높은 중복성으로 이어질 수 있으며, 이로 인하여 앙상블 모듈은 인체의 근육 라인(muscle line) 및 기타 낮은 수준의 특징(low-level features)을 별도로 학습해야 하기 때문이다. 반면, 도 2의 레벨 5와 같이 분기를 너무 늦게 시작하면 앙상블의 다양성이 크게 줄어들게 된다. 최적의 분기 수준(branch level)은 이러한 극단 사이에 있다고 추정할 수 있다.

[0073] 한편, 공유 레이어(shared layer)는 서브시퀀스 분기(subsequence branch)에서 얻은 정보를 토대로 공간적 특성(spatial features) 또는 추상적 감정의 특징(features)을 학습할 수 있다.

[0075] **<로컬 인식 모듈(LPM)>**

[0076] 포괄적인 얼굴 정보를 학습하여 더 높은 분류 정확도를 얻으려면 단순히 얼굴 표정의 전체 정보를 학습하는 것만으로는 충분하지 않을 수 있다. 얼굴 특징과 근육 활동은 우리가 특정 감정을 표현하는 동안 특정 패턴을 따를 수 있다. 본 발명은 로컬 인식 모듈(Local Perception Module, LPM)(102)을 설정하여 영역 세분화를 통해

과적합(overfitting) 문제를 어느 정도 완화할 수 있다.

[0077] 눈, 코, 입은 감정 표현에 가장 많이 기여하는 얼굴 부위로, 해당 영역을 관심 영역(ROI)으로 자르면, 전체 이미지의 중요하지 않은 부분으로 인해 발생하는 FER의 간섭을 줄일 수 있을 뿐만 아니라 작업에 민감한 주요 영역을 활용할 수 있다. 따라서 이러한 주요 지역을 잘라 전체 정보의 보완 역할을 할 수 있는 로컬 세부 정보를 얻을 수 있다.

[0078] 도 3은 얼굴의 영역 분해로, ROI의 각 점은 특정 좌표에 해당한다. 먼저 알고리즘을 사용하여 데이터 세트에서 얼굴을 감지하고 정렬하고, 각 왼쪽 눈의 ROI, 오른쪽 눈의 ROI 및 입과 코의 ROI를 포함할 수 있다.

[0079] 다음으로 인풋 얼굴을 두 정점 $(x_{\min}, y_{\min})$, $(x_{\max}, y_{\max})$의 좌표를 기반으로 세 개의 ROI 영역으로 자를 수 있다. 인풋 얼굴 이미지를 자른 후, 이러한 주요 ROI 영역은 LPM의 입력으로 균일하게 96 x 96 크기로 조정될 수 있다.

[0080] LPM의 구조는 GEMSR(104)의 단일 분기와 동일한 구조를 활용하여 LPM이 약분류기(weak classifier)로 동작할 수 있도록 한다. 그렇게 함으로써 앙상블 학습은 모든 약분류기 사이에서 일관성을 유지할 수 있으므로 모든 약분류기를 강분류기(strong classifier)로 통합할 수 있다.

[0081] 상기 약분류기는 이전 분류기가 잘못 분류한 샘플의 가중치를 적절하게 바꾸어 이전 분류기의 단점을 보완할 수 있고, 상기 강분류기는 개별 약 분류기들에 각 가중치를 적용하여 조합할 수 있다.

[0083] **<랭크 어텐션 유닛(Rank Attention Units, RAU 및 복수 보팅(plurality voting)>**

[0084] 도 4는, 왼쪽은 기존 분류 네트워크 스키마(CLS, classification token)와 오른쪽은 RAU(Rank Attention Unit)를 나타낸다.

[0085] 상기 RAU는 SE(Squeeze-and-Excitation) 레이어에서 영감을 받은 것으로 두 가지 요구사항을 충족해야 하는데, 첫째로, 선형 레이어 간의 비선형 상호 작용을 학습할 수 있는 충분한 유연성이 있어야 하고, 두번째로 각각의 분류 결과가 강조될 수 있도록 하여 모든 분기에 대한 배제가 없는 관계를 보장해야 한다.

[0086] 상기 RAU는 처음에 각 분기에서 얻은 분류 결과에 가중치를 두어 최종 선형 레이어와 동일한 크기의 1차원 또는 다차원 벡터를 구하고, 이러한 벡터는 결과값을 평가한 점수로 간주할 수 있다.

[0087] 그러한 유닛은 각 점수를 상응하는 초기 분류 결과값에 적용하고, 이는 각 분기(branch)에 가중치를 부여할 수 있다.

[0088] 한편, 일반적으로 앙상블의 판단 단계에서는 복수 보팅, 단순 평균 및 가중 평균의 세 가지의 널리 사용되는 규칙이 적용될 수 있다. 앙상블이란 여러 종류의 알고리즘을 사용하여 결합함으로써 보다 정확한 예측을 도출하는 기법이고, 보팅(voting)이란 여러 종류의 알고리즘을 사용한 각각의 결과에 대해 투표를 통해 최종 결과를 예측하는 방식이다.

[0089] 본 발명은 GEMSR의 정보를 효과적으로 통합하기 위해 복수 보팅(voting) 방식을 채택할 수 있는데, 로컬 인식 모듈을 활용하면 3개의 주요 얼굴 영역이 GEMSR에 통합되어 개별 학습기 간에 어느 정도 차이가 발생할 수 있으므로, 의사 결정 방법의 중요성을 고려하면서 모든 모듈에 가중치를 할당하여 각 분기의 분류 정확도를 검증할 수 있다.

[0090] 1차 앙상블 규칙 하에서 분류 결과의 최종 퓨전 방법은 복수 보팅(voting)일 수 있다.

[0091] 출력 Z(= [z1, z2,...zi])가 있는 분기 m(=1,...,M)이 다음을 위한 가중 벡터 S(= [s1, s2,...si])를 제공한다고 가정한다. 판단 단계에서 이러한 앙상블은 다음 [수학식 1]로 나타낼 수 있다.

수학식 1

[0092] $$S = F_{rank}(Z, W) = \sigma\left(W_2 \delta\left(W_1 Z\right)\right)$$

[0093] $W_1, W_2 \in \mathbb{R}^{C\times C}, \delta, \sigma$ 는 RELU 와 Sigmoid 함수를 의미한다.

[0094] 동시에 m번째 분기의 초기 분류 결과인 $Z^m = [z_1^m, z_2^m, \ldots z_i^m]$ 은 다음과 같이 주어진다.

수학식 2

[0095] $$Z^m = F_{CLS}(Z, W) = WZ$$

[0096] 이러한 매개 변수를 얻기위해 각 분기들에 $\tilde{Z}^m\ (= [\tilde{z}_1^m, \tilde{z}_2^m, \ldots \tilde{z}_i^m])$ 와 같은 가중치를 부여할 수 있다.

수학식 3

[0097] $$\tilde{z}_i^m = F_{\text{scale}}(z_i^m, s_i) = s_i \cdot z_i^m$$

[0099] 결국 복수의 보팅을 통해 결과를 얻고 이는 다음 [수학식 4]와 같이 표현될 수 있다. 여기서 Z_{final}의 값은 최종 퓨전 결과를 의미하고 c_i 는 주어진 감정 레이블을 의미한다.

수학식 4

[0100] $$Z_{\text{final}} = c_{\arg_i \max} \sum_{m=1}^{M} \tilde{Z}^m$$

[0101] 제안된 RAU가 실용적으로 사용되기위해 [수학식 5]와 같이 정의된 적합한 결합 교차 엔트로피 손실 함수를 최소화하여 종단 간 방식(end-to-end manner)으로 학습할 수 있다.

수학식 5

[0102] $$L_{GLFEN} = \sum_b \sum_i L\left[P\left(f\left(x_i\right) = y_i \mid x_i, \theta_b\right), y_i\right]$$

[0103] 여기서 b는 분기(branch)의 인덱스, x_i, y_i는 트레이닝 세트의 샘플을 반영하고 θ_b는 앙상블을 구성하는 컨볼루션 레이어의 매개변수를 나타낼 수 있다.

[0105] **< GLFEN 분석 >**

[0106] **ESR(Ensemble with Shared Representation)과의 비교**

[0107] ESR과 GLFEN 사이에는 여전히 주요한 차이점이 있다. ESR은 GEMSR의 일부로만 사용될 수 있다.

[0108] 첫째, 입력 이미지만 사용하는 ESR과 비교하여 GLFEN은 더 정확한 FER 예측을 달성하기 위해 주요 얼굴 영역에

서 로컬 특징을 추가로 추출할 수 있다.

[0109] 둘째, GLFEN은 글로벌 정보와 로컬 정보를 결합하여 정보의 보완성을 보장하고 앙상블의 다양성을 높일 수 있다.

[0110] 셋째, 어텐션 모듈을 사용함으로써 가려지거나 어수선한 얼굴을 인식하는 데 도움이 될 수 있다. 반면 ESR은 어텐션 모듈을 고려하지 않기 때문에 실제 또는 가려진 이미지를 처리하는 동안 성능이 좋지 않다. 본 발명은 RAU(Rank Attention Unit)를 사용하여 의사 결정 수준에서 신뢰성이나 중요성이 있는 의미 있는 가중치를 학습하므로 위의 문제를 완화할 수 있다.

[0112] **중복 분석**

[0113] 실시간 HCI(Human-Computer Interaction)를 수용하기 위해 네트워크의 매개변수와 계산 비용은 핵심 문제이다.

[0114] 본 발명은 ESR(Ensembles with Shared Representation)을 기반으로 GEMSR을 구성하는데, GEMSR의 도움으로 앙상블 네트워크의 공유 레이어가 매개변수와 계산 비용을 줄이는 데 도움이 될 수 있다.

[0115] 반면에 LPM의 네트워크 가중치는 매개 변수를 줄이기 위해 세 가지 ROI에서 공유된다.

[0116] GEMSR에서는 처음 몇 개의 공유 레이어를 적용하여 중복성을 줄이고 전체 네트워크의 학습 능력을 보장한다. 전통적인 앙상블 네트워크의 경우 분기가 확장됨에 따라 매개변수가 커지므로, 공유 레이어를 활용하면 GEMSR-4의 감소되는 매개변수는 거의 1/3 수준이다.

[0117] 매개변수 감소 비율은 분기 수가 증가함에 따라 증가할 것이다. 또한 공유 레이어에 의해 앙상블 크기도 줄었다. 결과적으로 전통적인 앙상블 네트워크에 비해 계산 비용이 줄어들 것이다.

[0118] LPM의 공유 가중치 입력 정보가 불완전하기 때문에 얼굴의 세 가지 주요 ROI에서 로컬 정보를 학습하기 위해 LPM을 채택했다. 그러나 학습을 위해 세 개의 개별 분기를 사용하면 네트워크 매개변수의 수가 두 배가 된다. 보완 정보를 활용하고 동시에 네트워크의 적당한 크기를 유지하기 위해 세 개의 들어오는 주요 ROI가 LPM에서 동일한 가중치를 공유한다.

[0120] **<다양한 데이터 셋에 대한 GLFEN 학습>**

[0121] 본 발명의 일 실시예로, in-the-lab dataset으로 훈련하고 검증할 수 있는데, 그 중 CK+ 및 JAFFE의 경우 데이터가 불충분한 반면에 고품질 및 선명한 얼굴 이미지를 제공하기 때문에 성능에 대한 증명을 만족스럽게 진행할 수 있다.

[0122] 본 발명의 일 실시예로, CK+ 및 JAFFE의 평가에서 LPM에서 5개의 컨볼루션 레이어를 사용하고, 각 레이어 뒤에는 각 배치별 평균 및 분산을 이용해 정규화하는 배치정규화(Batch Norm) 레이어와 경사함수(ReLU)가 뒤따를 수 있다.

[0123] 본 발명의 일 실시예로, 대규모 데이터셋 FER+ 및 RAF-DB의 특징 표현 능력을 향상시키기 위해 컨볼루션 레이어의 수를 9개로 늘릴 수 있다. 입력 로컬 영역이 동일한 LPM 가중치를 공유하도록 하여 네트워크 매개변수의 수를 크게 줄일 수 있다.

[0124] 알고리즘 1은 학습 GLFEN 알고리즘을 보여준다.

Algorithm 1 Training GLFEN.

Input: Incoming facial images and key regions data $\boldsymbol{D}$, the current branch index $\boldsymbol{m}$, the saved network $\boldsymbol{GLFEN_{m-1}}$ (if exist);
Output: Accuracy on training and validation data;
Divide the incoming data $\boldsymbol{D}$ into the validation data $\boldsymbol{D_e}$ and training data $\boldsymbol{D_r}$;
Set up the hyper-parameters, such as learning rate, optimal parameters ;
Initial the sharing layers of GEMSR and LPM with random parameters;
for $\boldsymbol{m}$ *ranges from* **0** *to* ***maximum branch idex*** **do**
  if $\boldsymbol{m = 0}$ **then**
    Initial the ensemble branch with random parameters;
  else
    Initial the branches with the saved network $\boldsymbol{GLFEN_{m-1}}$;
  end
  Add a new branch to the network ;
  Randomly sample a subset $\mathbf{D}^{\sim}$ from training data $\boldsymbol{D_r}$;
  for *each batch in the subset* $\mathbf{D}^{\sim}$ **do**
    Initialize the loss function $\boldsymbol{GLFEN_m}$ to **0**;
    Perform the forward phase;
    for *each existed branch* $\boldsymbol{Branch_m}$ *in the ensemble* **do**
      Compute the current loss $\boldsymbol{L_b^m}$ achieved by each branch $\boldsymbol{Branch_m}$;
      Add the current loss to $\boldsymbol{GLFEN_m}$;
    end
    Perform the backpropagation algorithm;
    Optimize the $\boldsymbol{GLFEN_m}$ with the forzen layers;
  end
  Apply the trained network on validation data $\boldsymbol{D_e}$ with the above training method;
  Save the current training network to $\boldsymbol{GLFEN_m}$;
end

[0125]

[0126] 본 발명의 일 실시예로, CK+ 데이터 셋을 대상으로 전처리 단계를 거친 후에 얼굴 이미지를 대상자 순서에 따라 10개의 폴드로 구분하고, 각 폴드에는 평균 12개의 물체와 131개의 얼굴 이미지가 채워질 수 있다.

[0127] 상기 알고리즘 1에 보고된 방법론에 따라 10개의 테스트를 수행했다.

[0128] GEMSR-4에 새로운 컨볼루션 분기를 추가한 후 GEMSR-4 학습률을 0(즉, 고정 레이어)으로 조정하면 만족스러운 분류 결과를 얻는 것으로 입증되었다.

[0129] 원래 앙상블 모듈이 완전한 기능을 학습했으며 과적합 문제로 이어지는 데 더 이상 학습이 필요하지 않기 때문에 새로 추가된 LPM은 원래 프레임워크에서 학습률 0.02로 미세 조정하는 것만으로 최적화되었다.

[0130] 여전히 큰 학습률을 유지하면 원래 모델이 학습한 내용이 망가질 가능성이 높다. 또한 SGD 방법에서 모멘텀 팩터 0.9를 사용하여 학습률이 0.5배(250 epoch마다) 감소한다. 마찬가지로 밝기, 대비 변경, 수평 뒤집기, 이미지 회전(30도) 및 크기 재조정을 포함한 과적합 문제를 줄이기 위해 데이터 확대도 무작위로 적용되었다.

[0131] 도 7은 분기 수준이 증가함에 따라 CK+ 데이터 세트의 원래 네트워크 성능과 비교하여 본 발명의 네트워크의 평균 정확도 사이의 관계를 보여준다. 이로부터 분기 수준 3이 분기 수준 2보다 낫다는 것을 알 수 있다. 분기 레벨 4의 성능이 좋지 않은 것으로 보아 너무 늦게 분기를 시작하면 다양성이 부족할 수 있다.

[0132] 표 1은 은 CK+에서 네트워크 크기, 테스트 정확도(%) 및 실행 시간(ms)을 비교한 것이다.

표 1

Method	Network size	Accuracy	Running time (ms)
Single Network	131.21k	85.5 ± 3.5%	0.1835
Traditional Ensemble	524.83k	89.2 ± 1.2%	0.2248
ESR-4 Lv.3 (Baseline) [42]	355.10k	89.4 ± 2.2%	0.1921
GLFEN-4 Lv.3	487.03k	**92.5 ± 1.9%**	0.2216

[0133]

[0134] 표 1에서는, 테스트 정확도 측면에서 single 네트워크의 정확도는 다른 세 가지 앙상블 방법보다 훨씬 낮아 앙상블 방법이 매우 효과적임을 나타내고, 나머지 세 가지 앙상블을 비교할 때 GLFEN-4 Lv.3의 테스트 정확도가 TE 및 ESR-4 Lv.3보다 우수한 것을 확인할 수 있다.

[0135] 표 1의 실행 시간 평가에 대해 단일 NVIDIA 2080 GPU를 사용하여 CK+ 데이터 세트에서 PyTorch로 네트워크를 테스트한다. 결과는 제안하는 방법이 평균 영상 처리 시간이 0.22ms로 실시간 구현(프레임 속도 24fps)에 적용할 수 있음을 나타낸다.

[0136] 도 8은 10개의 폴드를 만들어 교차 검증(10-fold Cross Validation)하는 것으로 네트워크와 기준선(baseline) 사이의 각 폴드(fold)의 평균 정확도(%)를 비교한 것이다.

[0137] 도 8은 10겹(fold) 교차 검증에서 GLFEN-4 Lv.3과 ESR-4 Lv.3 사이의 각 폴드(fold)의 예측 정확도를 나타낸다. GLFEN-4 Lv.3의 테스트 정확도가 각 폴드에서 ESR-4 Lv.3을 초과함을 알 수 있다. 이는 LPM과 RAU의 도입이 단순히 컨볼루션 레이어를 쌓는 것이 아니라 우수한 분류 결과를 얻을 수 있다는 실질적인 의의가 있다.

[0138] 네트워크 크기 측면에서 GLFEN-4 Lv.3은 중간 크기(487K)이다. 기준선과 비교하여 LPM의 구조는 단일 네트워크와 동일한 크기이므로 네트워크 매개변수의 증가는 약 131K에 불과함을 확인할 수 있다.

[0139] 표 2에서는, 8가지 방법에 대한 CK+ 정확도(%) 비교를 나타낸다.

표 2

Method	Avg Acc (8 class)
KNN-Chi-Square [28]	**94.18%**
GLFEN-4	92.54%
SVM-Chi-Square [28]	90.48%
RBF-Chi-Square [28]	90.38%
VGG-16 (Fine-Tune) [8]	89.90%
DSAE [50]	89.84%
ESR-4 (Baseline) [42]	89.43%
VGG-16 (From scratch) [8]	88.70%

[0140]

[0141] 표 2에서는, GLFEN-4와 다른 최첨단 방법인 KNN-Chi-Square, RBF-Chi-Square, SVM-Chi-Square, VGG- 16, DSAE, ESR-4 등 구체적인 비교 결과를 나타내고 있고, 네트워크 GLFEN-4는 기준 성능을 89.4%에서 92.5%로 상승시켰다.

[0142] KNN-Chi-Square가 가장 높은 정확도를 나타냈지만, 학습 대상 샘플의 클래스 불균형을 잘 처리할 수 없는 KNN(최근접 이웃) 알고리즘을 채택했고, 본 발명은 불균형 훈련 샘플을 사용하여 데이터 세트(예를 들면, FER+ 데이터 세트)에서 비슷한 성능을 달성할 수 있다.

[0143] 표 3에서는, JAFFE 데이터 세트의 평균 정확도(%) 비교를 나타낸다.

표 3

Method	Avg Acc (7 class)
GLFEN-4	**96.26%**
AutoFER [1]	94.97%
HDNN [15]	94.91%
STRNN [53]	94.89%
EMS-CNN [17]	92.85%
AFER [48]	92.69%
ESR-4 (Baseline) [42]	92.48%
GSP-KNN [29]	88.57%

[0144]

[0145] 상기 알고리즘 1의 훈련 단계에 따라 GLEFN-4의 분류 결과는 표 3에 나와 있다. 표 3에서는 가장 진보된 방법인 AutoFER, HDNN와 GLEFN-4의 성능을 비교하는데, GLEFN-4가 96.26%의 유리한 FER 정확도를 달성한다는 것을 확인할 수 있다. AutoFER과 비교하여 GLFEN-4는 정확도를 약 1.3% 향상시킨다. GLFEN-4의 정확도는 ESR-4(Baseline)에 비해 3.8% 우수하다.

[0146] 표 4에서는, 알고리즘 1로 모든 분기를 훈련시킨 후, FER+의 정확도(%) 비교를 나타낸다.

표 4

Method	Acc
GLFEN-9 (Fine-tune)	**88.77 ± 0.1%**
GLFEN-9 (From scratch)	88.64 ± 0.2%
SCN-ResNet18 [46]	88.01%
ResNet+VGG [13]	87.4%
ESR-9 (Baseline) [42]	87.15 ± 0.1%
SHCNN [30]	86.54%
VGG16-PLD [2]	84.99 ± 0.4%
TFE-JL [22]	84.3%
ResNet18 + FC [22]	83.4%

[0147]

[0148] FER+에서 학습하면서 네트워크의 효율성을 확인하기 위해 특별히 다음과 같은 전략을 사용할 수 있다. FER+의 사진은 모두 회색조 이미지이고 크기는 48X48이나, GLFEN-9의 컨볼루션 레이어와 일관성을 유지하기 위해 크기를 96X96 픽셀로 조정할 수 있다. SGD 방법을 사용하고, 모멘텀, 기본 학습률은 두 훈련 전략 모두에 대해 각각 0.9, 0.1로 설정될 수 있다.

[0149] GEMSR-9의 빠른 수렴으로 인해 학습률 감소의 승수(10 epoch마다)는 0.75로 증가했다. 학습을 위해 추가된 브랜치도 0.02의 기본 학습률로 지속적으로 업그레이드되었다.

[0150] GLFEN-9(From scratch)는 다른 데이터 세트에 대한 교차 훈련 전략 없이도 최첨단 방법보다 우수한 성능에 도달할 수 있으며, 이는 본 발명이 강력한 일반화 능력을 가지고 있음을 나타낸다.

[0151] GLFEN 9(Fine-tune)의 평균 테스트 정확도는 88.77%이며 표준 편차는 0.1%로 비교적 작다. 이전에 학습된 기능 중 일부가 관련된 시각적 인식 작업 간에 전송될 수 있으므로, 사전에 학습된 네트워크를 사용하면 처음부터 학습된 것보다 속도를 높이고 더 나은 예측을 달성할 수 있다.

[0152] Baseline과 비교하여 GLFEN-9(Fine-tune)가 1.62% 더 우수하다. 각 학습에 대한 샘플을 선택하고 대규모 데이터 세트에서 만족스러운 결과를 얻음으로써 네트워크가 더 강건하다고 설명할 수 있다.

[0153] 도 9는 FER+에 대한 앙상블 예측의 정규화된 컨퓨전 매트릭스(confusion matrix)를 나타낸다. baseline ESR-9(Fine-tune)와 비교하여 본 발명은 각각 31.25%의 경멸 샘플과 61.67%의 공포 샘플을 올바르게 인식했다. 반면 베이스라인은 각각 20%와 45.3%만 인식했다.

[0154] 또한 행복, 슬픔, 놀라움과 같은 다른 세 범주의 정확도는 모두 기준선을 어느 정도 초과한다. 본 발명은 56.25%의 경멸 샘플과 36%의 혐오 샘플을 중립 및 분노로 잘못 분류하는데, 경멸 및 혐오 범주에는 151개 및 157개 샘플만 있어 불균형 문제가 발생할 수 있다.

[0155] 사람들은 같은 표정을 보아도 복합적인 감정을 느낄 것이고, 이는 감정 인식의 주관적 문제로 이어질 수 있다. 이 편향 문제를 완화하기 위해 적절한 임계값에서 훈련 샘플 수를 제어할 수 있다.

[0156] 모든 감정 범주에는 앙상블의 다양성과 전반적인 정확도를 보장하기 위해 각 범주에 대한 상대적 최적 임계값인 FER+에서 선택된 최대 5000개의 훈련 샘플이 포함될 수 있다.

[0157] 표 5에서는, RAF-DB 데이터셋에 대한 평균 정확도의 정확도(%) 비교를 나타낸다.

표 5

Method	Avg Acc
GLFEN-9 (Fine-tune)	**84.72%**
DLP-CNN [54]	<u>84.13%</u>
pACNN [23]	83.27%
ESR-9 (Baseline) [42]	82.21%
ResNet-PL [35]	81.97%
GAN-Inpainting [49]	81.87%
VGG16 [39]	80.96%
ResiDen [16]	76.54%

[0158]

[0159] RAF-DB 데이터 세트의 많은 얼굴 이미지는 머리카락, 안경, 스카프, 호흡 마스크, 손, 팔, 음식 등과 같은 사실

적인 폐색에 의해 부분적으로 가려질 수 있다.

[0160] 도 10에는 위와 같은 가려진 샘플이 묘사되어 있다.

[0161] 표 5에서는 본 발명의 제안된 GLFEN-9가 비슷한 성능을 달성할 수 있음을 알 수 있다.

[0162] 본 발명은 지역 정보를 전역 정보의 보충 역할을 할 수 있는 지역 인식 모듈(LPM)을 활용하므로, 하나의 주요 얼굴 영역의 일부가 가려지면 다른 주요 영역의 국부적 특징도 추출 및 분류할 수 있다.

[0163] 본 발명은 RAU를 적용하여 각 분기의 분류 결과에 가중치를 할당할 수 있다. 예를 들어, 하나의 주요 얼굴 영역이 가려지면 로컬 분기의 출력에 상대적으로 낮은 가중치가 할당되어 최종 분류 결과에 미치는 영향을 약화시킬 수 있다. 따라서 얼굴이 부분적으로 가려지면 오분류를 어느 정도 줄일 수 있다.

[0164] 도 11은 RAF-DB에 대한 앙상블 예측의 정규화된 컨퓨전 매트릭스(confusion matrix)를 나타낸다. 데이터 세트 전체에 걸친 인식 성능의 일관성을 보여주기 위해 RAF-DB 데이터 세트에 대한 GLFEN-9(Fine-tune)의 혼동 행렬도 도 11에 표시된다.

[0165] 도 9의 결과와 비교하여 공포 및 혐오 샘플의 인식 정확도는 두 데이터 세트에서 상대적으로 낮다. 또한 22.35%의 공포 샘플과 19.88%의 혐오 샘플은 각각 놀라움과 분노 감정으로 오분류되어 FER+ 데이터 세트의 결과와 유사하다.

[0166] 인식 성능은 일반적으로 이 두 데이터 세트에서 일관되지만 RAF-DB(59.92%)에 대한 혐오 샘플의 인식 정확도는 FER+(48%) 데이터 세트보다 높다. RAF-DB 데이터셋의 감정 혐오 훈련 샘플이 FER+ 데이터셋보다 많기 때문에 혐오 샘플의 불균형 문제는 RAF-DB 데이터셋에서 어느 정도 완화될 수 있다.

[0167] 표 6은 성능에 영향을 주는 요인(features)을 제거하거나 변형하면서 각 요인이 최종 성능에 미치는 영향을 비교하는 어블레이션 실험(ablation experiment)의 결과를 보여준다.

표 6

[0168]

	Acc 0	Acc 1	Acc 2	Acc 3	Acc 4	Acc 5
LPM		✓			✓	✓
RAU-S			✓		✓	
RAU-M				✓		✓
CK+	89.4 ± 2.2%	91.4 ± 3.5%	91.4 ± 2.4%	91.0 ± 2.9%	**92.5 ± 1.9%**	92.7 ± 3.7%
JAFFE	92.5 ± 4.6%	94.3 ± 4.9%	94.8 ± 4.4%	94.9 ± 4.8%	**96.3 ± 4.1%**	96.3 ± 4.4%
FER+	87.15 ± 0.07%	88.56 ± 0.12%	88.60 ± 0.04%	88.58 ± 0.02%	**88.77 ± 0.05%**	88.75 ± 0.13%
RAF-DB	82.21 ± 0.09%	83.88 ± 0.15%	83.83 ± 0.05%	83.81 ± 0.08%	**84.72 ± 0.07%**	84.66 ± 0.18%

[0169] GLFEN에는 LPM(Local Perception Module)과 RAU(Rank Attention Unit)를 포함한 두 가지 추가 구성 요소가 포함되어 있어서, 기여도를 개별적으로 확인하기 위해 일련의 어블레이션 실험을 수행할 수 있다.

[0170] LPM의 경우, Baseline은 공유 레이어와 4개의 앙상블 분기만 포함하는 GEMSR-4를 참조하므로, 각 학습자는 5개의 컨볼루션 레이어를 갖을 수 있다.

[0171] LPM의 구조는 GEMSR-4의 기본 학습기와 일치하는데, 표 6의 결과에서와 같이 GEMSR-4의 성능은 LPM에 의해 더욱 향상되어 LPM이 없는 실험보다 최소 1.4% 이상 향상될 수 있다. 전체 네트워크가 LPM을 활용하여 완전한 정보를 학습하기 때문이다.

[0172] RAU-S는 주의 가중치 벡터가 분류기의 각 출력에 대해 단일 차원이므로 복수의 보팅 전에 각 출력의 감정 예측에 동일한 가중치를 곱한 것을 의미한다.

[0173] RAU-M은 주의 가중치 벡터가 다차원이므로 모든 감정 예측에 서로 다른 가중치가 곱해짐을 의미한다.

[0174] RAU-S와 RAU-M 모두 RAU-S와 RAU-M이 없는 것보다 더 나은 성능을 얻는 것을 알 수 있다.

[0175] 이는 RAU가 성능을 결정하는 데 중요한 역할을 한다는 것을 나타낸다. 또한 RAU-M과 비교하여 모든 데이터셋에

대한 RAU-S의 분류 결과는 유사하지만 표준편차가 더 낮다.

[0176] 학습 샘플이 부족한 경우. 시각적 증거로 RAU의 효율성을 추가로 확인하기 위해 분류 작업에서 차별 영역을 로컬화할 수 있는 Grad-CAM을 사용하여 클래스 활성화 매핑의 시각화를 제공할 수 있다.

[0177] 도 12에서는, 다른 얼굴 표정의 시각적 결과를 나타낸다. 얼굴의 활성화된 영역이 강조 표시되고 빨간색 영역은 최종 분류 결과에 가장 기여하는 특징을 나타낸다. RAU가 없는 방법과 비교할 때 활성화 맵은 분노, 혐오, 행복 및 슬픔과 같은 대부분의 표현에 대해 넓은 범위의 영역을 다루는 경향이 있다. 반면에 RAU의 도움으로 차별적인 영역을 강조할 수 있다.

[0178] 예를 들어 분노 표현의 경우 RAU를 사용한 방법은 얼굴의 가장 기여도가 높은 영역(예: 눈썹, 입, 눈)에 집중할 수 있다. 이 때문에 RAU는 얼굴의 분류에 유리한 영역에 집중하도록 네트워크를 안내하여 유리한 예측 정확도를 달성할 수 있다.

부호의 설명

[0179] 100 : GLFEN 모듈

102 : 로컬 인식 모듈

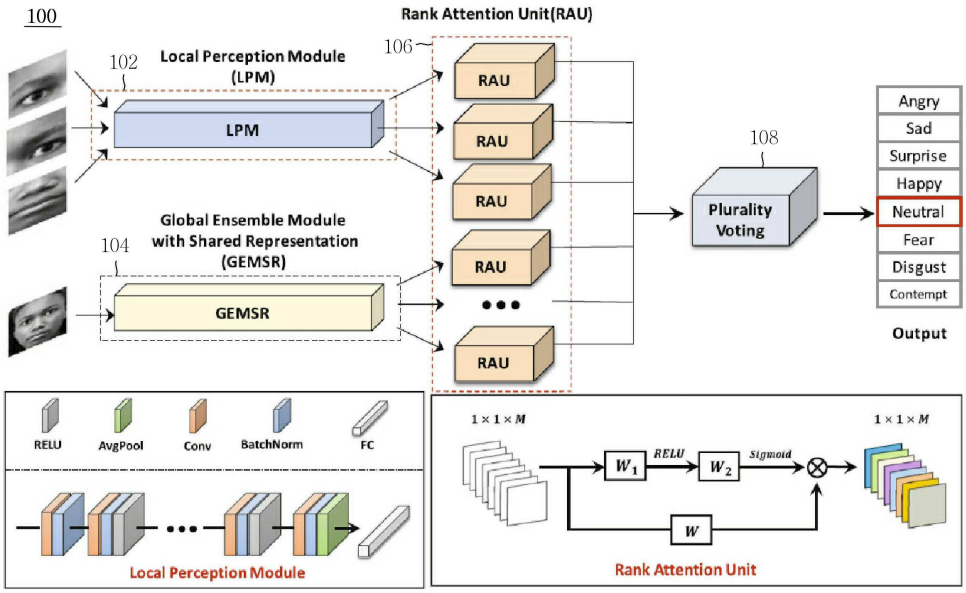
104 : 공유 표현 글로벌 앙상블 모듈

106 : 랭크 어텐션 유닛 모듈

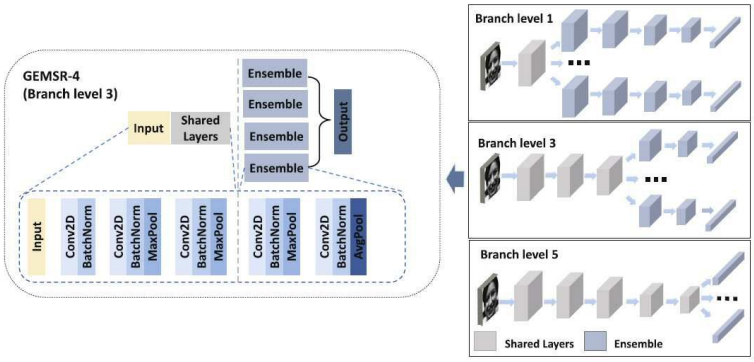
108 : 복수 보팅 모듈

도면

도면1



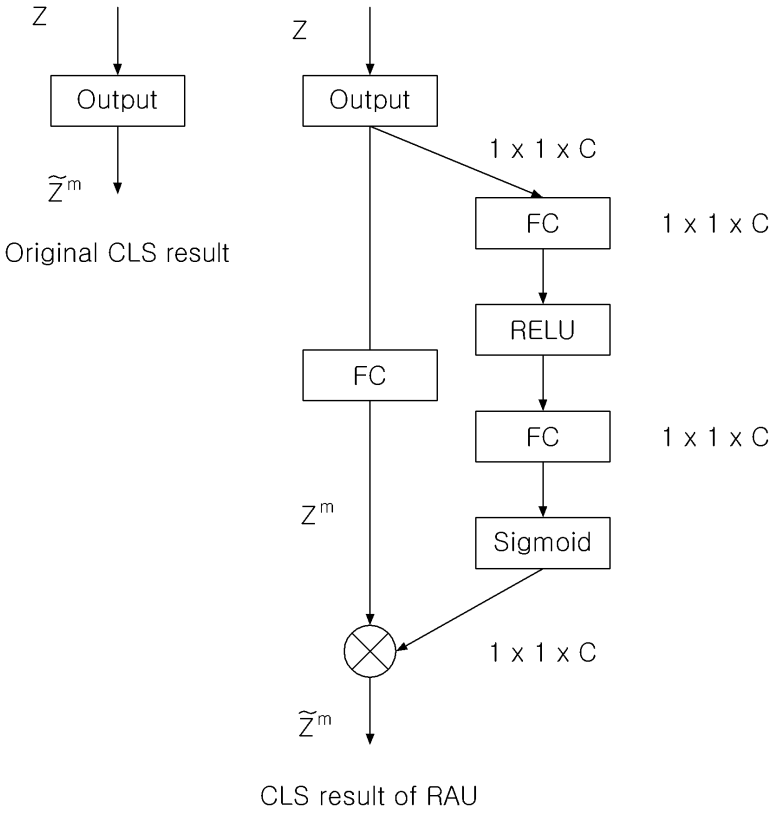
도면2


GEMSR-4
(Branch level 3)
Input
Shared Layers
Ensemble
Ensemble
Ensemble
Ensemble
Output
Input
Conv2D
BatchNorm
Conv2D
BatchNorm
MaxPool
Conv2D
BatchNorm
MaxPool
Conv2D
BatchNorm
MaxPool
Conv2D
BatchNorm
AvgPool
Branch level 1
Branch level 3
Branch level 5
Shared Layers
Ensemble

도면3


ROI of the left eye
ROI of the left eye
ROI of the mouth and nose

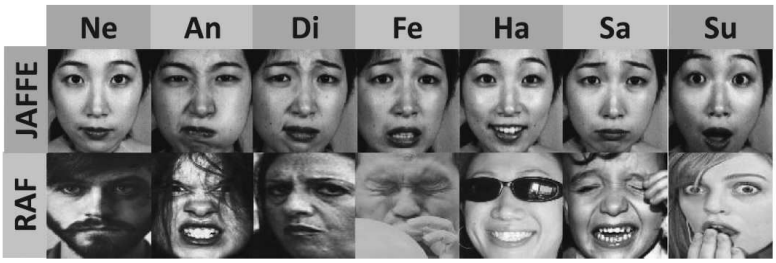
도면4


Z
Output
$\widetilde{Z}^m$
Original CLS result
Z
Output
1 x 1 x C
FC
1 x 1 x C
RELU
FC
FC
1 x 1 x C
Z^m
Sigmoid
1 x 1 x C
$\widetilde{Z}^m$
CLS result of RAU

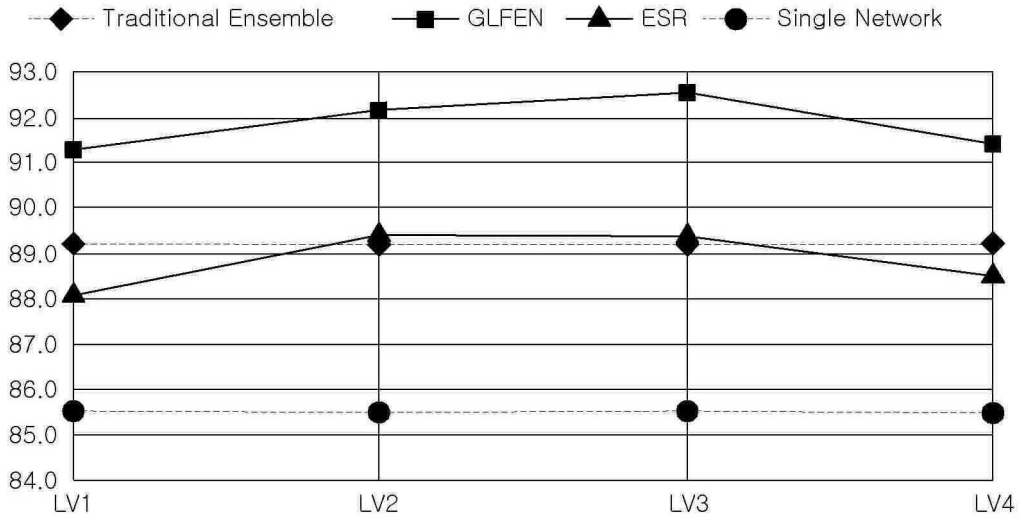
도면5


Ne
An
Co
Di
Fe
Ha
Sa
Su
CK+
FER+

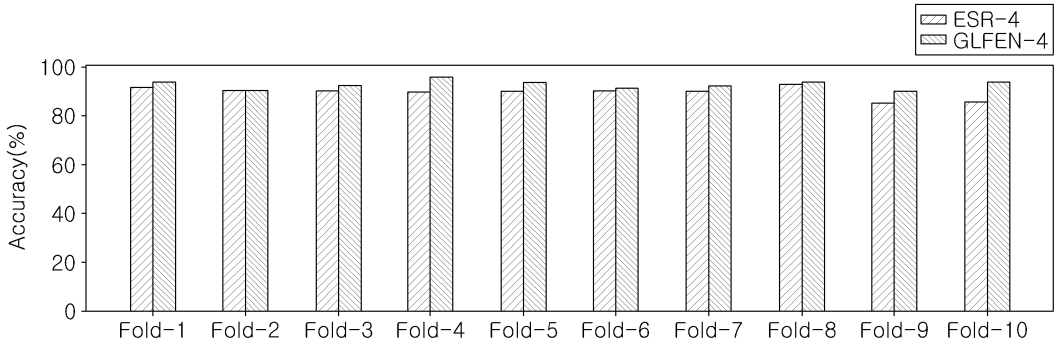
도면6


Ne
An
Di
Fe
Ha
Sa
Su
JAFFE
RAF

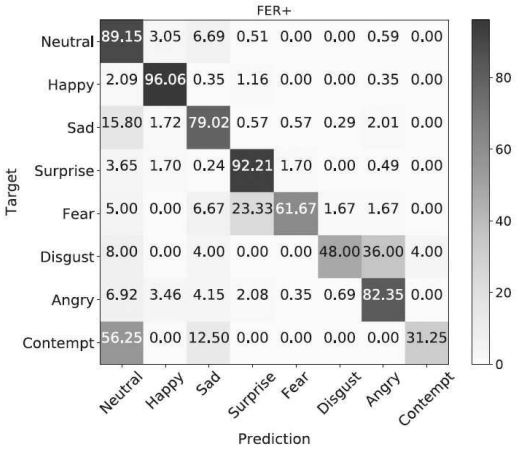
도면7


Traditional Ensemble
GLFEN
ESR
Single Network
93.0
92.0
91.0
90.0
89.0
88.0
87.0
86.0
85.0
84.0
LV1
LV2
LV3
LV4

도면8

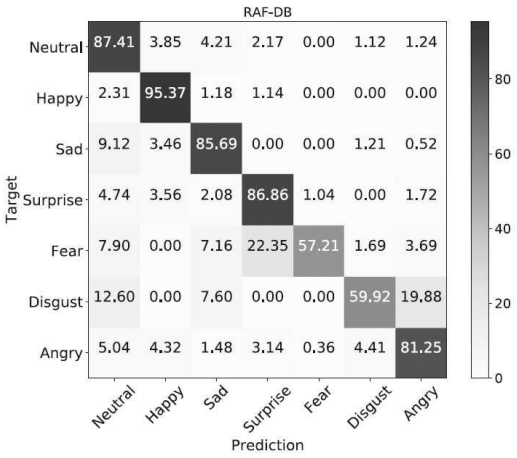

ESR-4
GLFEN-4
Accuracy(%)
100
80
60
40
20
0
Fold-1
Fold-2
Fold-3
Fold-4
Fold-5
Fold-6
Fold-7
Fold-8
Fold-9
Fold-10

도면9

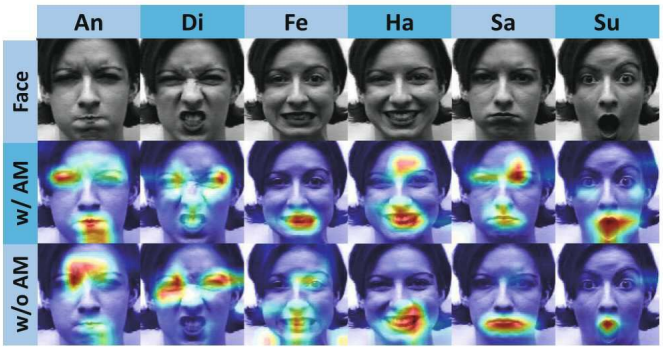

FER+
Target
Neutral 89.15 3.05 6.69 0.51 0.00 0.00 0.59 0.00
Happy 2.09 96.06 0.35 1.16 0.00 0.00 0.35 0.00
Sad 15.80 1.72 79.02 0.57 0.57 0.29 2.01 0.00
Surprise 3.65 1.70 0.24 92.21 1.70 0.00 0.49 0.00
Fear 5.00 0.00 6.67 23.33 61.67 1.67 1.67 0.00
Disgust 8.00 0.00 4.00 0.00 0.00 48.00 36.00 4.00
Angry 6.92 3.46 4.15 2.08 0.35 0.69 82.35 0.00
Contempt 56.25 0.00 12.50 0.00 0.00 0.00 0.00 31.25
Neutral
Happy
Sad
Surprise
Fear
Disgust
Angry
Contempt
Prediction
80
60
40
20
0

도면10

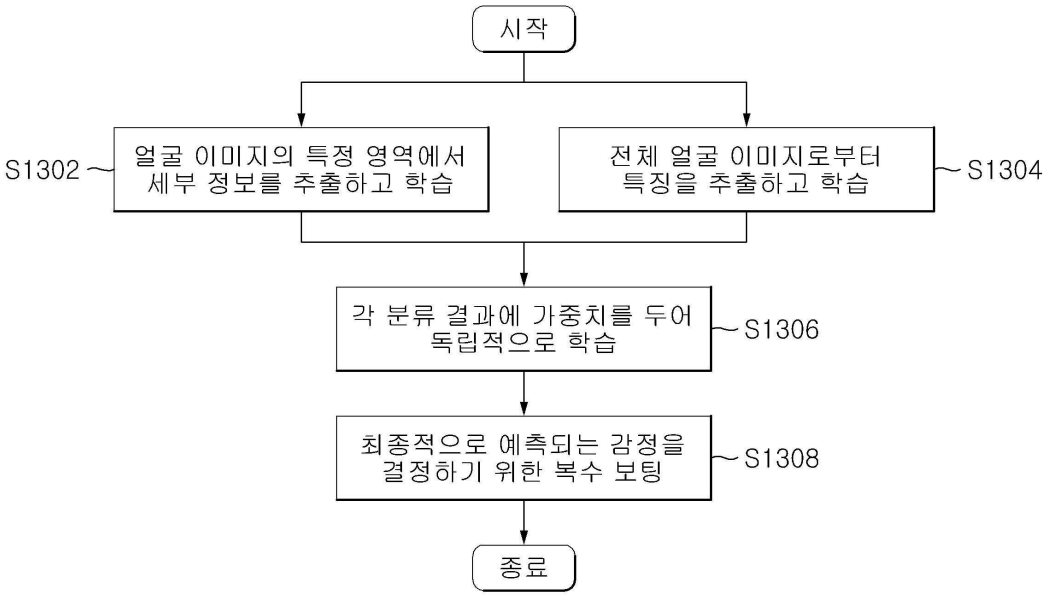
도면11



도면12



도면13


시작
S1302
얼굴 이미지의 특정 영역에서 세부 정보를 추출하고 학습
전체 얼굴 이미지로부터 특징을 추출하고 학습
S1304
각 분류 결과에 가중치를 두어 독립적으로 학습
S1306
최종적으로 예측되는 감정을 결정하기 위한 복수 보팅
S1308
종료