Journal of KIIT. Vol. 21, No. 5, pp. 39-47, May 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.5.39

얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달 감정인식 시스템

조찬영*, 정현준**

Multimodal Emotion Recognition System using Face Images and Multidimensional Emotion-based Text

Chanyoung Jo*, Hyunjun Jung**

이 논문은 정부(과학기술정보통신부)의 재원으로 한국 연구재단의 지원을 받아 수행된 연구임 (No. NRF-2022R1G1A1008493), 본 연구는 환경부/한국환경산업기술원 지중환경오염위해관리기술개발사업 (2022002450002)으로 수행되고 있습니다

요 약

기존 컴퓨터 분야의 감정인식은 한 가지의 형태와 사람의 통상적인 감정을 바탕으로 연구가 진행되었다. 이는 사람의 감정을 정확히 인식하는 데 문제점이 있다. 사람의 감정을 정확히 인식하려면 다양한 정보가 필요하며, 연속적인 공간에서 감정의 변화를 인식해야 한다. 본 논문에서는 이런 문제점을 해결하고자 다양한 형태와 감정을 다차원 공간으로 인식한 VAD 신호를 이용한 멀티모달 감정인식을 제안한다. 제안한 멀티모달의 구성은 얼굴 감정인식 CNN 모델, 텍스트 감정인식 BERT 모델을 결합했다. 제안 방법의 우수성을 입증하기 위해 단일 감정인식 모델과 기존 멀티모달을 비교했다. 비교결과 본 논문에서 제안한 멀티모달이 감정인식이 6% 성능 향상을 보였다는 점에서 정확한 감정인식을 했다는 기여점을 보였다.

Abstract

Emotional recognition in the existing computer field was conducted based on one form and normal human emotions. For accurate emotion recognition, various forms and emotions must be recognized on a continuous line. In this paper, to solve this problem, we propose multimodal emotion recognition through VAD that recognizes various forms and emotions as multidimensional spaces. The proposed multimodal configuration combines CNN and BERT models. To demonstrate the superiority of the proposed method, a single emotion recognition model and existing multimodal were compared. As a result of the comparison, the multimodal proposed in this paper showed a contribution to accurate emotional recognition in that emotional recognition showed improved 6% performance.

Keywords
multimodal, emotional recognition, VAD, CNN, BERT

* 군산대학교 소프트웨어학과 학사과정
- ORCID: https://orcid.org/0000-0002-5557-4090
** 군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-6717-1395

· Received: Feb. 24, 2023, Revised: Mar. 27, 2023, Accepted: Mar. 30, 2023
· Corresponding Author: Hyunjun Jung
Dept. of Software at Kunsan National University, 558, Daehak-ro, Kunsan-si, Jeollabuk-do, Republic of Korea
Tel.: +82-63-469-8917, Email: junghj85@kunsan.ac.kr

Ⅰ. 서 론

사람이 감정을 표현하는 매개체로는 음성, 표정, 자세가 있다. 말은 감정의 특성을 반영할 수 있는 매개변수를 포함하고 있기에 같은 문장이라도 듣는 사람에게 다르게 느껴지는 경우가 있다[1]. 표정도 감정의 중요한 외적 형태로서 특정한 감정 정보를 담고 있다[2]. 그렇기에 한 가지의 형태(Unimodal)만 가지고 감정을 인식한다는 것은 많은 어려움이 있어 다양한 형태(Multimodal)를 통해 감정을 인식해야 한다[3].

감정을 인식하는 것은 의사소통에서 매우 중요하다. 비즈니스 분야에서는 고객의 감정 상태를 조기에 예측하거나 인식하지 못하면 고객들이 무시당하는 느낌을 받게 된다는 연구가 있다[4]. 이처럼 감정은 심리학 분야에서 오랜 연구 대상이었으며, 컴퓨터 분야에서도 감정을 인식하려는 연구가 진행중이다[5]. 컴퓨터 분야의 기존 감정인식 연구는 감정의 상태를 인간의 통상적인 감정을 분류하여 모델을 학습한다[6]. 이는 인위적인 것으로 정확한 감정의 상태를 파악하는 것은 부족하며, 한 가지의 형태만 가지고 감정을 인식했다는 문제점이 있다.

최근에는 감정을 다차원 공간으로 인식하기 위한 연구[7]와 다양한 형태를 통해 감정을 인식하는 연구가 진행되었다[8]. 감정을 다차원 공간으로 인식하기 위한 연구는 감정을 다차원 공간에 표현하여 감정의 어떤 성분이 어느 정도 내포되어 있는지를 파악하기 위해 차원 축에 감정의 영역을 설정하여 감정을 분류한다. 다양한 형태를 통해 감정을 인식하는 연구는 하나의 형태의 데이터를 사용하는 것보다 다양한 형태를 가진 데이터를 사용하는 것이 모델의 성능을 높였다.

본 논문에서는 하나의 형태만 이용하여 사람의 감정을 인식하는 문제점과 감정을 다차원으로 인식하기 위해 얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달 감정인식 방법에 대하여 제안한다. 제안하는 시스템은 얼굴 정보를 이용한 감정인식 모델과 음성정보를 텍스트로 변환하여 텍스트를 이용한 다차원 감정인식 모델로 구성된다.

본 논문에 구성은 다음과 같다. 2장에서는 관련 연구로 멀티모달을 이용한 감정인식과 다차원 감정인식 연구에 관해 기술하고, 3장에서는 본 논문이 제안하는 얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달 감정인식 방법에 관해 기술한다. 4장에서는 제안한 방법을 실험 및 평가하고, 마지막으로 5장에서는 결론에 관해 기술한다.

Ⅱ. 관련 연구

본 장에서는 멀티모달 감정인식과 다차원 감정인식 연구에 관한 동향을 설명한다.

2.1 멀티모달 감정인식

기존의 감정인식 모델은 텍스트나 이미지를 이해하는 하나의 형태에 중점을 두었다. 하지만 정확한 감정을 인식하기 위해서는 다양한 형태의 데이터를 통해 감정을 인식해야 한다. 최근에는 다양한 형태의 데이터를 이용한 감정인식 연구가 진행되었다[9]-[11].

[9]는 얼굴 영상과 음성을 이용한 감정인식 멀티모달을 제안했다. 얼굴 영상을 이용한 감정인식은 패턴인식 분야에서 사용되고 있는 방법인 2D-PCA(2D Principal Component Analysis)를 적용한 모듈로 구성했다. 음성을 이용한 감정인식은 MFCC(Mel-Frequency Cepstral Coefficient) 기법을 통해 특징을 추출했다. 멀티모달 감정인식 시스템의 성능 향상을 위해 특징 단계, 유사도 단계, 결정 단계로 구분하였으며 얼굴과 음성의 인식 결과를 유사도 단계에서 결합하여 최종 감정을 분류했다. 본 논문에서도 얼굴 정보와 텍스트 정보에 가중치를 적용하여 모델의 정확도를 비교한다.

[10]은 한국어 영상 데이터 감정 분류를 위한 멀티모달을 제안했다. 동영상과 음성에 특화된 딥러닝 모델을 구상했으며 동영상 모델은 3D 합성곱 신경망(3D convolution neural network)을 사용하여 특징을 추출하며 음성 모델은 MFCC 기법을 통해 음성 데이터를 시각화하여 표현할 수 있는 스펙트로그램 이미지로 만들었다. 두 가지 특징 추출방법을 통해 나온 특징 벡터를 병합하여 소프트맥스(Softmax) 함수를 통해 가장 확률이 높은 감정을 예측했다.

[11]은 음성과 텍스트를 이용한 멀티모달을 제안했다. 음성 감정인식은 ResNet(Residual Network) 모델을 사용했으며, 텍스트 감정인식 모델은 BERT(Bidirectional Encoder Representations from Transformers)를 사용했다. 멀티모달의 감정 분류는 음성 및 텍스트 모델에서 얻은 점수의 가중 평균을 사용했다. 가중 평균은 각 모델의 소프트맥스 계층(Softmax layer) 출력에 자연로그를 적용하여 계산하며 가장 높은 점수를 받은 감정 범주를 선택한다.

2.2 다차원 감정인식

사람의 감정을 다차원으로 인식하기 위해 최근에는 매우 세밀한 연속적인 감정 정보인 VAD (Valence-Arousal-Dominance)를 이용한 감정인식 연구가 진행되었다[12]-[14]. 그림 1은 Russell이 제안한 다차원 감정 이론의 차원 축인 Valence, Arousal, Dominance에 매칭된 감정을 표현한 것이다[12]. 매칭된 감정들은 고유의 VAD 벡터값을 갖고 있다.

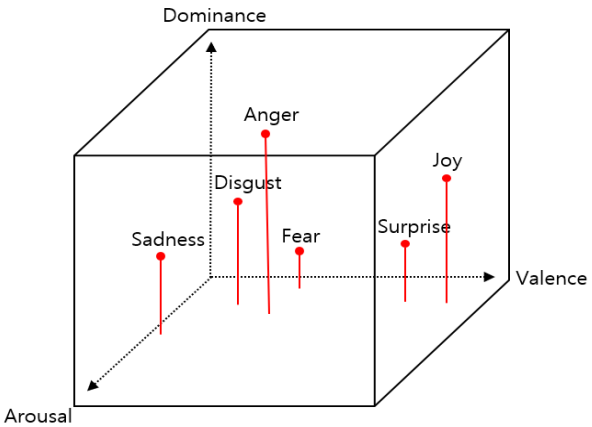

그림 1. VAD 감정 차원
Fig. 1. VAD emotional dimensions

표 1은 VAD 차원 축의 의미이다. Valence는 감정이 어느 정도로 유쾌한지(Positive) 혹은 불쾌한지(Negative)를 나타낸 지표이며, Arousal은 감정으로 인해서 얼마나 신체적인 흥분 상태(Excited)인지 진정상태(Calm)인지, Dominance는 감정으로 인해 통제력을 제어할 수 있는지(Controllable) 제어할 수 없는지(Uncontrollable)를 나타낸다.

표 1. VAD 차원의 의미
Table 1. Meaning of the VAD dimension

Dimension	Meaning
Valence	Positive - Negative
Arousal	Excited - Calm
Dominance	Controllable - Uncontrollable

[13]은 상담 챗봇에서 내담자가 작성한 문장에서 감정을 정확하게 파악하기 위해 상담 챗봇의 다차원 감정인식 모델을 제안했다. 기존 텍스트 감정인식은 대부분 감정 키워드나 어휘를 분석하여 한가지 감정으로 분류하였다는 한계점이 있다. 해당 연구에서는 감정 키워드로 감정을 분석하는 것이 아닌 VAD 데이터를 통해 모델을 학습했다. Word2Vec 임베딩을 통해 단어를 벡터화시켰으며, LSTM(Long Short-Term Memory)의 단점인 정보 손실 문제와 기울기 소실 문제를 해결하기 위해 Attention으로 모델을 구성했다.

[14]는 VAD 신호를 이용하여 감정 분류를 수행하는 기술에 대해 다룬다. 기계 학습 기술인 SVM(Support Vector Machine), k-NN(k-Nearest Neighbor), MLP(Multilayer Perceptron) 알고리즘을 사용하여 VAD 신호에서 감정 분류를 수행하는 실험을 진행하였고, 각 알고리즘의 성능을 비교 분석했다. 실험 결과 MLP 알고리즘이 가장 높은 분류 성능을 보였다. 해당 연구는 VAD 신호를 이용한 감정 분류 기술이 가능하다는 것을 입증했다. 또한, 이 기술은 인간 감정인식이 어려운 상황에서 유용하게 활용될 수 있다는 가능성을 제시했다.

III. 얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달 감정인식 시스템

본 장에서는 본 논문이 제안하는 얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달의 구조와 동작 방식에 관해 설명한다.

3.1 시스템 구조

본 논문에서 제안하는 얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달 시스템은 그림 2와 같다. 멀티모달의 전체적인 구조는 얼굴 정보 처리 과정(Facial information processing)과 음성 정보 처리 과정(Voice information processing), 최종 감정 도출(Emotion recognition result)로 구성된다. 얼굴 정보 처리 과정에서 얼굴 감정인식을 위해 인간의 시각 처리 과정을 모방한 딥러닝 모델인 CNN(Convolution Neural Network) 모델을 사용한다[15]. CNN 모델을 통해 인식된 얼굴 감정은 소프트맥스 과정을 통해 모든 감정의 확률이 계산된다. 음성 정보 처리 과정에서는 음성을 텍스트로 바꾸는 STT(Speech-to-Text) 기법을 사용한다. STT 기법은 사람의 음성 인터페이스를 통해 텍스트 데이터를 추출한다. 음성인식은 음성에 대한 파형을 분석하는 것으로 시작하여 입력 음성이 들어오면 음성 전처리 단계를 거친다. 전처리 과정에서 특징 벡터 열 반환하여 패턴을 인식하고 언어처리 과정을 통해 텍스트 데이터로 변환이 된다. 변환된 텍스트는 감정인식 모델의 입력값으로 들어간다. 텍스트 감정인식 모델은 사전학습된 자연어처리 모델인 BERT를 사용한다[16]. BERT 모델을 통과한 텍스트는 회귀 과정을 통해 예측된 VAD 값이 나오며 VAD 감정 지표 벡터값을 통해 모든 감정의 코사인 거리를 계산한다. 최종 감정 도출은 각 모델의 출력으로 나온 감정을 수식을 적용한 우선순위를 계산하며, 우선순위가 높은 감정을 최종 감정으로 선택한다.

3.2 얼굴 감정인식 모델

본 절에서는 얼굴 정보의 감정을 인식하는 모델에 대하여 설명한다.

영상에서 사용자의 감정을 추출하기 위해 사용되는 얼굴 감정인식 모델의 구조는 그림 3과 같다. 입력 이미지를 받으며 합성곱(Convolution2D) 과정에서 필터를 통해 입력 이미지의 가중치를 계산하며 배치(Batch)마다 평균과 분산을 활용하여 데이터의 정규화 과정인 배치 정규화(BatchNormalization)를 거친다. 두 번의 과정을 거치면 입력 신호의 총합을 출력 신호로 변환하기 위해 활성화 함수 ReLU(Rectified Linear Unit)를 사용한다. ReLu 함수를 통과한 값들은 평균 풀링(AveragePooling2D) 과정을 통해 특징을 추출한다. 모델의 과적합 방지를 위해 드롭아웃(Dropout)은 0.5로 설정한다. 4번의 과정을 반복하면 합성곱 과정, 배치 정규화, 합성곱 과정 순서로 특징을 추출하며 글로벌 평균 풀링(Global AveragePooling) 과정을 통해 이미지의 크기를 줄이고 소프트맥스 함수를 통해 모든 감정에 대한 확률을 계산한다.

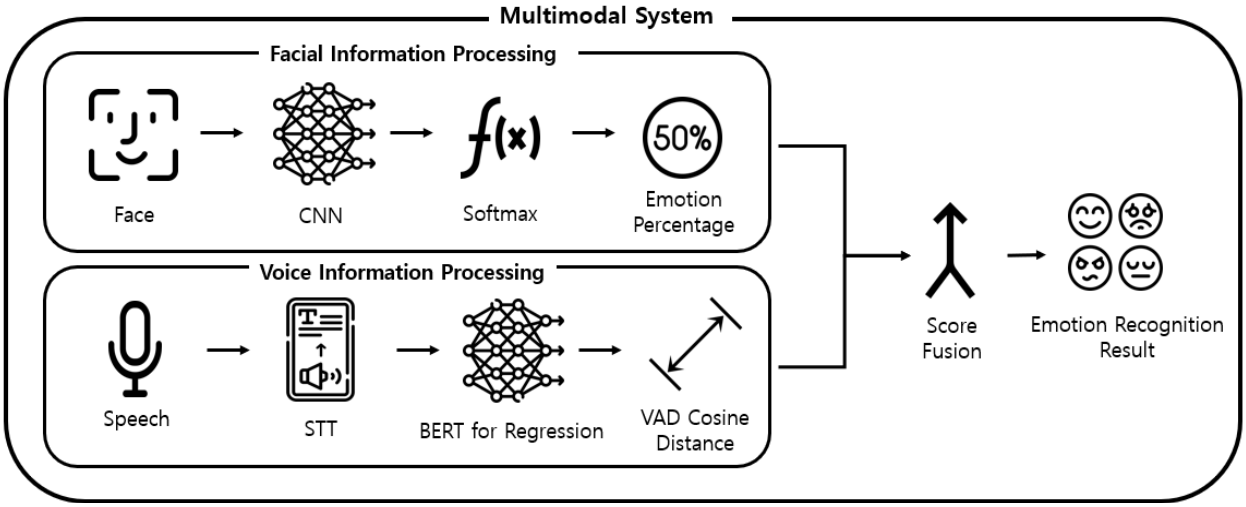

그림 2. 제안 멀티모달 시스템
Fig. 2. Proposed multimodal system

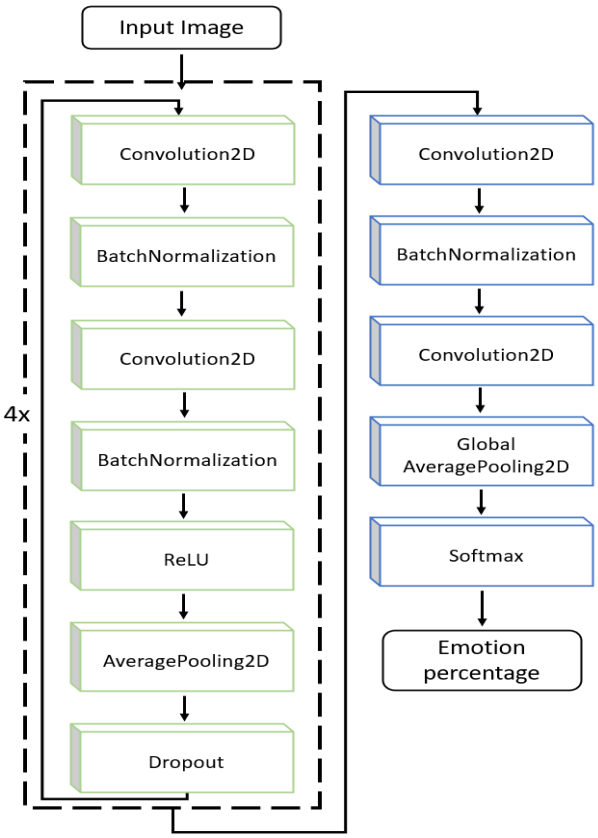

그림 3. 얼굴 감정인식 모델 구조
Fig. 3. Facial emotion recognition model structure

3.3 다차원 감정 기반의 텍스트 감정인식 모델

본 절에서는 다차원 감정 기반의 텍스트 감정인식 모델에 대하여 설명한다.

영상에서 받아 온 음성정보는 파이썬(Python)에서 제공하는 음성인식 패키지 SpeechRecognition를 통해 음성을 텍스트로 변환해주는 작업을 수행한다[17]. 다차원 감정을 인식 모델의 구조는 그림 4과 같다. STT를 통해 자연어로 변환된 데이터는 토큰화 과정(Tokenization)을 거쳐 자연어처리 모델인 BERT의 입력값으로 들어간다. 기존 BERT 모델은 분류의 문제를 해결하기 위해 모델이 구성되어 있지만, 본 논문에서는 다차원 감정의 VAD 값을 예측하기 위해 출력층에 회귀계층(Regression Layer)을 추가하였다. VAD 값은 3개를 예측해야 하기에 마지막 출력층은 3개로 설정했다. 예측된 VAD 값은 VAD 감정 지표 값을 통해 모든 감정의 코사인 거리를 계산한다.

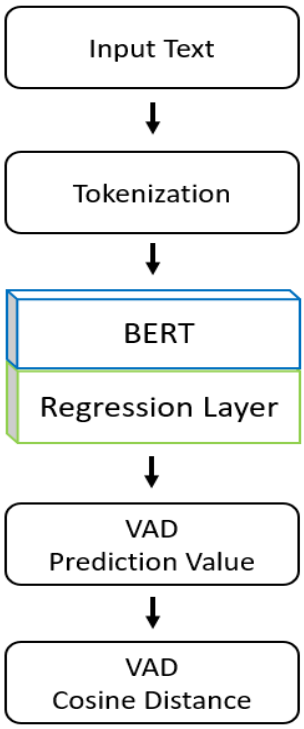

그림 4. 텍스트 감정인식 모델 구조
Fig. 4. Textual emotion recognition model structure

3.4 멀티모달 감정인식

본 절에서는 멀티모달의 최종 감정 결정 방법에 관해 설명한다.

모델의 감정인식 분류는 분노(Anger), 혐오(Disgust), 두려움(Fear), 기쁨(Joy), 슬픔(Sadness), 놀라움(Surprise)으로 총 6가지이다. 얼굴 감정인식 모델에서는 소프트맥스 함수를 통해 모든 표정에 대한 확률을 구하여 감정의 우선순위를 정한다. 다차원 감정인식 모델은 BERT 모델을 통해 나온 VAD 값과 VAD 감정 지표 값을 이용한다. 표 2는 6가지 감정에 대해 NRC-VAD 사전에 매칭한 VAD 값을 나타낸다[18]. 각 감정은 VAD 차원에 매칭한 고유의 Valence, Arousal, Dominance의 벡터값이 존재한다. ‘Anger’는 (0.167, 0.865, 0,657), ‘Disgust’는 (0.052, 0.775, 0,317), ‘Fear’는 (0.073, 0.84, 0,293), ‘Joy’는 (0.98, 0.824, 0,794), ‘Sadness’는 (0.052, 0.288, 0,164), ‘Surprise’는 (0.875, 0.875, 0,562)의 VAD 벡터값을 갖는다.

표 2. VAD 감정 벡터값
Table 2. VAD emotion vector value

Emotion	Valence	Arousal	Dominance
Anger	0.167	0.865	0.657
Disgust	0.052	0.775	0.317
Fear	0.073	0.84	0.293
Joy	0.98	0.824	0.794
Sadness	0.052	0.288	0.164
Surprise	0.875	0.875	0.562

VAD 감정 분류 알고리즘은 Algorithm 1과 같다. Line 1은 다차원 감정인식 모델이 출력한 VAD 값이다. Line 2~7은 각 감정의 고유의 VAD 벡터값이다. Line 8은 다차원 감정인식 모델이 출력한 VAD 값과 각 감정의 고유의 VAD 벡터값의 거리를 계산하는 함수이다. Line 9는 계산된 각 감정의 거리 값이 들어간다. Line 10~15는 각 감정에 대한 코사인 거리를 계산한다. 식 (1)은 코사인 거리의 수식을 나타낸다. Line 16은 계산된 거리가 가장 가까운 감정을 반환한다.

Algorithm 1 : VAD emotion classification

```
VAD = [ Input Text VAD Value ]
Anger = [ 0.167, 0.865, 0.657 ]
Disgust = [ 0.052, 0.775, 0.317 ]
Fear = [ 0.073, 0.84, 0.293 ]
Joy = [ 0.98, 0.824, 0.794 ]
Sadness = [ 0.052, 0.288, 0.164 ]
Surprise = [ 0.875, 0.875, 0.562 ]
def vad_cos(VAD):
 dist = { Emotion Distance }
 dist['Anger'] = distance.cos(VAD, Anger)
 dist['Disgust'] = distance.cos(VAD, Disgust)
 dist['Fear'] = distance.cos(VAD, Fear)
 dist['Joy'] = distance.cos(VAD, Joy)
 dist['Sadness'] = distance.cos(VAD, Sadness)
 dist['Surprise'] = distance.cos(VAD, Surprise)
 return min(dist)
```

$$distance.\cos(X, Y) = 1 - \frac{XY}{\| X \|_2 \| Y \|_2} \quad (1)$$

식 (2)는 최종 감정을 선택하는 수식이다. 두 개의 감정 정보를 융합하기 위해 얼굴 정보의 가중치를 부여한다. E_t와 E_s는 각각 얼굴과 텍스트에 대한 각 감정에 대한 우선순위를 의미한다. p는 얼굴 감정에 부여하는 가중치이며, 텍스트 감정에 대한 가중치는 $(1-p)$로 모든 가중치의 합은 1이 된다. E_m은 두 개의 감정을 계산한 우선순위 값들이며 우선순위가 제일 높은 감정을 최종 감정으로 선택한다.

$$\max(E_m) = pE_t + (1-p)E_s,\ 0 \le p \le 1 \quad (2)$$

Ⅳ. 실험 및 평가

본 장에서는 해당 논문에서 제안하는 멀티모달의 성능을 평가한다.

4.1 실험 방법

모델 학습을 위한 실험 환경은 표 3과 같다. 학습에 사용한 소프트웨어는 Python 3.8, Keras 2.3.1, PyTorch 1.13.1 버전을 사용했다. 얼굴 감정인식 모델의 학습 데이터는 48x48 그레이스케일 이미지로 구성된 얼굴 이미지를 7가지 범주 중 하나로 분류한 FER-2013 데이터셋을 사용했다[19]. 해당 데이터는 본 논문에서 제안하는 감정 분류의 맞게 6가지 감정으로 재분류를 했다. 다차원 감정인식 모델의 학습 데이터는 단어가 가지는 감성적인 특성을 파악하기 위해 영어, 한국어 등 여러 다국어 VAD 값이 있는 VAD-Lexicon 데이터셋을 사용했다[20]. 학습된 얼굴 감정인식의 모델의 정확도는 65%이며, 다차원 감정인식의 모델의 성능은 회귀 평가를 위한 지표인 MSE(Mean Squared Error)는 0.06이며 R2 score는 0.7이다.

표 3. 실험 환경
Table 3. Experimental environment

Component	Specification
CPU	Intel(R) Core(TM) i9-13900K
GPU	NVDIA GeForce RTX 4090
SSD	256GB
OS	Window 11
Software	Python 3.8 Keras 2.3.1 PyTorch 1.13.1

4.2 실험 결과 및 비교 평가

본 논문에서 제안하는 멀티모달의 성능 평가를 위해 단일 모델일 때 성능을 확인하며, [10]에서 제안한 Seokho 멀티모달 모델과의 성능 비교를 위해 [10]에서 사용한 감정 분류 데이터셋을 이용하여 평가한다[21]. [10]은 영상과 음성을 사용하여 비교하였으며 본 논문에서는 영상과 음성을 텍스트로 변환한 과정을 거쳐 비교 평가를 진행했다. 모델의 감정인식 정확도 평가를 위해 혼돈 행렬을 이용하여 평가했다. 표 4는 개별 감정인식 모델을 사용하여 혼돈 행렬 형태로 도시화한 결과이다. 얼굴 감정인식에 대한 정확도는 65%, 텍스트 감정인식에 대한 정확도는 68%를 보였다.

표 5는 식 (2)를 적용한 얼굴 감정인식 모델과 다차원 감정인식 모델의 가중치의 따른 정확도를 보여준다. 본 논문에서 제안하는 멀티모달 시스템의 인식률은 얼굴 가중치가 0.6일 때 74%라는 성능을 보였다. 이는 얼굴만 이용한 결과 9%, 텍스트만 이용한 결과 6%의 향상된 정확도를 보였다.

표 6은 본 논문에서 제안한 멀티모달과 Seokho 멀티모달의 성능을 비교한 결과이며, 정확도는 0.5, F1-score는 0.4가 향상되었다.

표 5. 가중치의 따른 멀티모달 정확도
Table 5. Multimodal accuracy according to weight

Weight of face	Weight of text	Recognition accuracy
0.0	1.0	68%
0.1	0.9	69%
0.2	0.8	70%
0.3	0.7	71%
0.4	0.6	72%
0.5	0.5	73%
0.6	**0.4**	**74%**
0.7	0.3	73%
0.8	0.2	71%
0.9	0.1	68%
1.0	0.0	65%

표 6. 성능 비교결과
Table 6. Performance comparison results

Model	Accuracy	F1-score
Seokho[10]	0.69	0.70
Proposed system	0.74	0.74

V. 결 론

본 논문에서는 기존 컴퓨터 분야의 감정인식의 문제점인 한 가지의 형태와 인간의 통상적인 감정을 바탕으로 한 감정인식의 문제를 해결하기 위해

표 4. 개별 감정인식 모델 정확도
Table 4. Accuracy of individual emotion recognition models

Facial emotion model						
Emotion	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger	65%	4%	10%	0%	7%	14%
Disgust	17%	61%	10%	7%	0%	5%
Fear	8%	6%	63%	3%	15%	5%
Joy	6%	6%	0%	76%	7%	5%
Sadness	4%	0%	7%	4%	72%	13%
Surprise	5%	20%	3%	6%	10%	56%

Text emotion model						
Emotion	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger	59%	10%	21%	0%	4%	6%
Disgust	13%	68%	6%	8%	0%	5%
Fear	8%	5%	65%	2%	20%	0%
Joy	7%	5%	0%	75%	2%	11%
Sadness	0%	5%	12%	5%	73%	5%
Surprise	4%	14%	3%	9%	2%	68%

다양한 형태인 얼굴 영상과 음성 및 텍스트를 사용하였으며, 통상적인 감정 분류를 해결하고자 감정을 다차원 공간으로 인식한 VAD를 사용했다. 얼굴 감정인식 모델은 CNN 모델을 사용했으며, FER-2013 데이터셋으로 학습을 진행했다. 음성처리 및 텍스트 감정인식은 STT 방식을 통해 음성을 텍스트를 변환하여 VAD 데이터셋을 학습한 BERT 모델을 사용했다.

실험 평가에서 단일 감정인식 모델을 사용했을 때 보다 두 개의 모델을 융합한 멀티모달이 우수한 성능을 보였으며, 정확도는 74%를 보였다. 이는 얼굴 영상만 사용했을 때 보다 9%, 텍스트 정보만 사용했을 때 보다 6%의 향상된 결과이다. 기존 단일 감정 분류 멀티모달 비교 평가에서는 본 논문이 제안한 멀티모달이 6%의 향상된 정확도를 보였다.

이후 진행될 감정인식 분야 연구에서도 다차원 감정을 기반으로 한 멀티모달을 활용한다면 기존 멀티모달 대비 높은 정확도를 바탕으로 다양한 연구가 진행될 것으로 기대된다.

References

[1] J. J. Guyer, P. Briñol, T. I. Vaughan-Johnston, L. R. Fabrigar, L. Moreno, and R. E. Petty, "Paralinguistic Features Communicated through Voice can Affect Appraisals of Confidence and Evaluative Judgments", J Nonverbal Behav, Vol. 45, No. 4, pp. 479-504, Jul. 2021. https://doi.org/10.1007/s10919-021-00374-2.

[2] J. A. Russell, "Facial expressions of emotion: what lies beyond minimal universality?", Psychol Bull, Vol. 118, No. 3, pp. 379-391, Nov. 1995. https://doi.org/10.1037/0033-2909.118.3.379.

[3] Y. Park and I. Shoji, "Three-Year-Old Children Focus on Emotional Adjectives When Linguistic Context and Facial Expression were Not Congruent: The Priming Task Research", The Korean Society of Emotional and Behavioral Disorders, Vol. 33, No. 1, pp. 51-70, Mar. 2017. http://doi.org/10.33770/JEBD.33.1.3.

[4] P. Chamola and P. Tiwari, "Customer delight and mood states: an empirical analysis in Indian retail context", International Journal of Indian Culture and Business Management, Vol. 8, No. 4 pp. 543-554, Jun. 2014. https://doi.org/10.1504/IJICBM.2014.062482.

[5] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection", Artificial Intelligence Review, Vol. 53, pp. 2313-2339, Oct. 2020. doi:https://doi.org/10.1007/s10462-019-09770-z.

[6] C. Jo, D. Kim, A. Yang, and H. Jung, "Emotion expression technique of Virtual environment player using Emotion analysis model", Proc. of KIIT Conference, Jeju, Korea, pp. 248-250, Jun. 2022.

[7] S. Buechel and U. Hahn, "Emotion Analysis as a Regression Problem – Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation", ECAI'16: Proc. of the Twenty-second European Conference on Artificial Intelligence, pp. 1114-1122, Aug. 2016. https://doi.org/10.3233/978-1-61499-672-9-1114.

[8] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion", Information Fusion, Vol. 37, pp. 98-125, Sep. 2017. https://doi.org/10.1016/j.inffus.2017.02.003.

[9] H. Go, Y. T. Kim, and M. Chun, "A Multimodal Emotion Recognition using Face Image and Speech", Journal of the Korea Society of Digital Industry and Information Management, Vol. 8, No. 1, pp. 29-40, Mar. 2012.

[10] S. Moon and S. B. Kim, "Multimodal Deep Learning Model for Korean Video Sentiment Classification", Journal of the Autumn Conference of the Korean Industrial Engineering Association, pp. 2944-2955, Nov. 2020.

[11] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal Emotion Recognition using Transfer Learning from Speaker Recognition and

BERT-based models", arXiv:2202.08974, Feb. 2022. https://doi.org/10.48550/arXiv.2202.08974.

[12] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions", Journal of Research in Personality, Vol. 11, No. 3, pp. 273-294, Sep. 1977. https://doi.org/10.1016/0092-6566(77)90037-X.

[13] M. J. Lim, M. H. Yi, and J. H. Shin, "Multi-Dimensional Emotion Recognition Model of Counseling Chatbot", Smart Media Journal, Vol. 10, No. 4, pp. 21-27, Dec. 2021.

[14] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion Classification Based on Biophysical Signals and Machine Learning Techniques", Symmetry, Vol. 12, No. 1, pp. 21, Dec. 2019. https://doi.org/10.3390/sym12010021.

[15] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis", Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings, Edinburgh, UK, pp. 958-963, Aug. 2003. https://doi.org/10.1109/ICDAR.2003.1227801.

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805.

[17] D. Amos, "The Ultimate Guide To Speech Recognition With Python", https://realpython.com/python-speech-recognition [accessed: Feb. 22, 2023]

[18] S. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words", Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, Vol. 1, pp. 174-184, Jul. 2018. http://dx.doi.org/10.18653/v1/P18-1017.

[19] M. Sambare, "FER-2013 dataset", https://www.kaggle.com/datasets/msambare/fer2013 [accessed: Feb. 22, 2023]

[20] S. Mohammad, "The NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon", https://saifmohammad.com/WebPages/nrc-vad.html [accessed: Feb. 22, 2023]

[21] AIhub, "dataset for emotion classification", https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=259 [accessed: Feb. 22, 2023]

저자소개

조 찬 영 (Chanyoung Jo)

2020년 3월 ~ 현재 : 군산대학교 소프트웨어학과 학사과정
관심분야 : 자연어처리, 웹, 서버

정 현 준 (Hyeonjun Jung)

2008년 : 삼육대학교 컴퓨터과학과(학사)
2010년 : 숭실대학교 컴퓨터학과(공학석사)
2017년 : 고려대학교 컴퓨터·전파통신공학과(공학박사)
2017년 8월 ~ 2020년 8월 : 광주과학기술원 블록체인인터넷경제연구센터 연구원
2021년 ~ 현재 : 군산대학교 소프트웨어학과 교수
관심분야 : 블록체인, 데이터 사이언스, 센서 네트워크, 사물인터넷, 머신러닝