Journal of Korea Multimedia Society Vol. 20, No. 8, August 2017(pp. 1175-1186)
https://doi.org/10.9717/kmms.2017.20.8.1175

이미지 시퀀스 얼굴표정 기반 감정인식을 위한 가중 소프트 투표 분류 방법

김경태[†], 최재영[††]

Weighted Soft Voting Classification for Emotion Recognition from Facial Expressions on Image Sequences

Kyeong Tae Kim[†], Jae Young Choi[††]

ABSTRACT

Human emotion recognition is one of the promising applications in the era of artificial super intelligence. Thus far, facial expression traits are considered to be the most widely used information cues for realizing automated emotion recognition. This paper proposes a novel facial expression recognition (FER) method that works well for recognizing emotion from image sequences. To this end, we develop the so-called weighted soft voting classification (WSVC) algorithm. In the proposed WSVC, a number of classifiers are first constructed using different and multiple feature representations. In next, multiple classifiers are used for generating the recognition result (namely, soft voting) of each face image within a face sequence, yielding multiple soft voting outputs. Finally, these soft voting outputs are combined through using a weighted combination to decide the emotion class (e.g., anger) of a given face sequence. The weights for combination are effectively determined by measuring the quality of each face image, namely "peak expression intensity" and "frontal-pose degree". To test the proposed WSVC, CK+ FER database was used to perform extensive and comparative experimentations. The feasibility of our WSVC algorithm has been successfully demonstrated by comparing recently developed FER algorithms.

Key words: Emotion Recognition, Facial Expression Recognition, Weighted Soft Voting, Face Sequences, Combination Weights

1. 서 론

최근 자동화된 감정인식(emotion recognition) 기술은 인간-컴퓨터 상호 작용과 데이터를 이용한 애니메이션 등 여러 분야에 영향을 미치고 있다. 얼굴표정(facial expression)은 사람의 감정을 대변하는 뛰어난 특징을 가지며 감정상태를 효과적으로 전달하는 가장 강력하고 자연스러운 방법 중 하나이다[1, 40]. 또한 표정은 인간의 의사소통을 위한 주요 방법으로 알려져 있다. 지금까지 표정인식에 관한 많은 연구가 진행되었지만, 얼굴표정의 미묘함, 복잡성 및 가변성을 극복 할 수 있는 표정인식 알고리즘 개발은 여전히 미흡하다[2,3,4].

컴퓨터 시스템이 사람의 얼굴표정을 통해 감정상

※ Corresponding Author : Jae Young Choi, Address: (17035) 81 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Korea, TEL : +82-31-330-4906, FAX : +82-31-330-4906, E-mail : jychoi@hufs.ac.kr
Receipt date : Jul. 17, 2017, Approval date : Jul. 24, 2017
[†] Dept. of Computer and Electronic Systems Eng., Hankuk University of Foreign Studies
(E-mail : kyeongtae.kim@hufs.ac.kr)
[††] Dept. of Computer and Electronic Systems Eng., Hankuk University of Foreign Studies
※ This research was supported by Hankuk University of Foreign Studies Research Fund.
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01057420)

태를 인식하여 감정인식 결과를 제공할 수 있는 초지능(artificial super intelligence) 알고리즘들을 개발하고 실생활에 적용하려는 시도가 최근에 활발하게 이루어지고 있다[5]. 현재 많은 연구를 통해 개발된 감정인식 기술은 영화, 드라마, 게임 등 다양한 콘텐츠에 노출되는 사용자의 솔직한 평가나 반응을 알기 위한 시장조사 분야에서 많이 사용되고 있다. 기존 설문조사는 콘텐츠를 감상한 후에 평가 내리는 방식이라 감상자의 편향, 주관 등이 개입될 수 있다. 하지만 감정인식 분석기술을 활용하면 콘텐츠를 즐기는 도중에 사용자의 표정으로 평가해 객관적인 데이터를 파악할 수 있다. 최근 마이크로소프트(MS)가 개발한 감정인식 프로그램은 영상 속 얼굴을 인식해 감정을 읽는다. 감정을 분노·경멸·혐오·공포·행복·중립·슬픔·놀람 등 8가지로 세분화해 수치화된 데이터로 보여준다[6]. 이 기술을 이용하면 식당에서 음식을 먹는 손님 표정으로 그날의 음식, 서비스 등을 평가받고 개선하는 데 도움을 받을 수 있다. 미국의 스타트업 업체인 '어펙티바(Affectiva)'는 여러 사람의 표정을 카메라에 담은 뒤 이들의 감정을 범주화해 '표정 데이터베이스'를 구축했다. 이를 통해 코카콜라, 유니레버 등의 광고를 소비자에게 보여주고, 이들이 이에 반응하는 표정을 분석할 수 있는 소프트웨어를 개발했다. 이 소프트웨어는 기업의 마케팅을 위한 데이터로 활용되고 있다[7].

얼굴표정기반 감정인식의 초기 연구는 정적(정지상태, static state) 얼굴 영상에서 감정 분석에 초점을 맞추었다. 하지만 감정 분석이 동적 스트리밍(dynamic streaming)으로 바뀌면서 연구의 초점이 영상 시퀀스(스트림)를 통한 감정 인식으로 바뀌고 있는 추세다. 즉, 영상 시퀀스를 사용하는 얼굴영상 기반 감정인식은 자동 감정 인식 분야에서 비교적 새로운 연구 주제라 할 수 있다. 영상 시퀀스를 활용하는 기존 감정인식 기술들을 다음과 같이 분류 할 수 있다. 1) 키-프레임(key-frame) 기반 연구방법[8, 9,10], 2) 얼굴의 동적특징 기반 연구방법[11,12,13], 3) 시공간 기술자(spatio-temporal descriptor) 기반 연구방법. 첫 번째 기존 연구방법은 감정인식을 수행하기 위해 정지 영상의 정적특징을 기반으로 한다. 대부분의 방법들에 경우 키-프레임(key-frame)에서 선택된 대표적인 피크 표현 얼굴(peak expression faces)만이 얼굴인식에 사용된다. 첫 번째 연구방법의 한계는 피크 표현 얼굴(peak expression faces)을 자동으로 선택하는데 있어 정확성에 대한 보장이 어렵기 때문에 신뢰할 수 있는 얼굴감정인식 성능을 제공하지 못한다는 점이다. 두 번째 기존 연구방법은 영상 시퀀스에 내재된 얼굴영상의 역학정보를 활용하기 위해 영상 시퀀스에 포함된 모든 얼굴영상들이 시간적 순서와 함께 이용된다. 일반적으로 분노와 같은 기본적인 얼굴표정의 동역학 모델을 학습하기 위해 HMM(Hidden Markov Model)과 같은 확률 모델을 사용한다. 두 번째 연구방법에 문제점은 획득된 영상 시퀀스에 포함된 얼굴영상들이 시간적으로 연속적이지 않은 경우가 종종 발생하여 실질적인 응용에서 얼굴영상의 역학정보를 신뢰성 있게 모델링 할 수 없다는 점이다. 세 번째 기존 연구방법은 시공간 기술자를 이용하여 공간 및 시간적 식별 정보(spatial and temporal discriminative information)를 모두 활용하는 것으로 목표로 한다. 일반적인 방법으로는 세 개의 직교평면으로부터 추출된 LBP-TOP 특징정보[14,15]와 세 개의 직교평면으로부터 추출된 LPQ-TOP[15,16] 특징정보를 활용하는 방법이 있다. 그러나 시간에 따른 표정 변화의 다양한 특성들(예를 들어, 얼굴근육 전이 혹은 전이 유형)이 동적 특징 추출(dynamic feature extraction)에 부정적으로 영향을 미칠 수 있기 때문에, 이러한 연구방법들은 영상 시퀀스에서 차별적인 공간 및 시간 정보 추출하는데 용이하지 않을 수 있다.

본 논문에서는 얼굴표정에서 감정을 효과적으로 인식하기 위해 가중 소프트 투표 분류 알고리즘(weighted soft voting algorithm)을 제안한다. 제안하는 가중 소프트 투표 분류 알고리즘은 영상 시퀀스의 활용을 극대화하여 감정인식을 수행하는 것을 목표로 한다. 이를 위해 제안한 방법은 비디오 시퀀스 내에 존재하는 얼굴영상들을 서로 다른 특징표현들(feature representations)로 학습된 다중 분류기들에(multiple classifiers) 입력하고 획득한 표정인식 투표들(votes)을 효과적으로 결합하여 표정인식 성능을 최적화한다. 제안 방법의 기술적인 특징들은 다음과 같다.

• 입력된 얼굴 영상의 최종 감정인식 분류결과는 주어진 얼굴영상에서 추출된 특징표현들로 훈련된 다중 분류기들에 의해 결합된다. 이 아이디어의 핵심

은 다양하고 상이한 특징표현들이 표정인식을 위한 얼굴영상의 상이한 특성들을 내포하고 있어 일부 모호한 표정인식 특징정보를 명확한 인식이 가능하게 한다. 따라서 상이한 특징 표현으로 훈련된 다중 분류기들에서 수집된 투표 결과들을 이용하여 얼굴 표정의 복잡한 패턴을 정확하게 인식하기 위한 차별적인 정보들이 상호 보완적으로 작용한다는 점에서 표정인식 정확도(accuracy)를 높일 수 있다.

• 각각의 분류기가 불확실한 얼굴 영상 분류에 대해 동일하지 않은 성능을 나타내는 경우, 과반수 투표(majority voting)[17] 기반 분류 프레임워크에서 다른 가중치(weight)를 부여하여 얼굴감정인식 성능을 향상시킬 수 있다. 따라서, 정확한 인식을 위해 더욱 강력한 (또는 더 신뢰할 수 있는) 투표 결과들에 더 큰 가중치를 부여하기 위한 효과적인 가중치 결정(weight determination) 알고리즘을 개발하였다. 제안한 가중치 결정 알고리즘은 얼굴 영상에 대한 분류기 신뢰도와 현재 입력 얼굴 시퀀스들에 포함된 각 얼굴영상의 품질(quality)을 계산하고 분류기들의 투표들의 가중치들을 계산하는 데 사용된다. 이를 위해 얼굴 시퀀스의 품질 척도로서 '표정의 피크 표현 강도(peak expression intensity)'와 '정면자세 정도(frontal-pose degree)'를 정의하고 효과적으로 구한다.

비디오 표정인식에서 가장 많이 활용되는 데이터베이스인 CK+ DB[18]를 사용하여 체계적으로 제안 방법의 우수성을 검증하였다. CK +DB를 활용해서 평가되었던 기존 표정인식 성능들과 비교하여 제안한 방법은 향상된 감정인식 정확도를 획득하였다.

2. 개 요

Fig. 1은 얼굴 시퀀스(face sequence)를 입력으로 하여 표정인식을 수행하기 위해 제안하는 방법을 전체적으로 보여준다. Fig. 1에서 얼굴 시퀀스는 하나의 동일한 감정유형(예: 분노)을 가진 일련의 다중개의 얼굴영상들로 구성되어 있다고 가정한다. 또한 얼굴 시퀀스는 얼굴추적(face tracking)[19]과 얼굴검출(face detection)[20] 기술들을 사용하여 얼굴 영역을 포함하는 비디오 시퀀스에서 획득한 것이라 가정하자.

제안 방법에서는 입력 얼굴 시퀀스로부터 오정렬된(misaligned) 얼굴영상들 및 비얼굴영상(nonface images)들을 지원벡터기계(Support Vector Machine, SVM)[21] 분류기 기술을 통해 제거한다. 오정렬된 얼굴영상들 혹은 비얼굴영상들이 제거된 얼굴 시퀀스는 다중개의 특징표현들로 학습된 다중 분류기들에 각각 입력되어 얼굴 시퀀스에 포함된 각 얼굴영상에 대해 특정 감정유형(emotion class)에 대한 소프트 투표(soft voting) 값들을 얻는다. 또한, 얼굴 시퀀스에 포함된 얼굴영상들의 품질(quality)을 측정하기 위해 얼굴영상의 '피크표현 강도(peak expression intensity)'와 '정면자세 정도(frontal-pose degree)'를 계산한다. 얼굴 시퀀스 품질 값들은 가중치 결정 알고리즘에 적용하여 소프트 투표값에 부여될 가중치 값을 구한다. 제안하는 가중치 소프트 투표 알고리즘은 소프트 투표(soft voting)값들과 소프트 투표 값들에 부여된 가중치를 결합하여 입력 얼굴 시퀀스의 최종 감정부류를 결정한다.

3. 오정렬된 얼굴영상 및 비얼굴영상 제거 방법

실질적인 응용에서는 잘못된 얼굴 검출 및 추적 결과로 여러 개의 오정렬된 얼굴 영상 및 비얼굴영상들이 얼굴 시퀀스에 포함될 수 있다. 주어진 얼굴 시퀀스에 대해 올바른 얼굴 감정인식을 수행하기 위해 이러한 불필요한 얼굴 영상들은 감정인식을 수행하기 전에 제거가 되어야 한다. 이를 위해, SVM 분류기를 활용하여 불필요한 얼굴영상들과 올바르게 검출된 얼굴영상들을 구별하였다. 긍정 혹은 부정 부류(class)가 표기(labelling)된 훈련 샘플들을 $S=\{(x_i,l_i)\}_{i=1}^{L}, l_i\in\{-1,1\}$라 하자. 여기서 x_i는 얼굴 시퀀스에 포함된 i-번째 얼굴 영상을 나타낸다. 또한 Fig. 2에서와 같이 오정렬된 얼굴영상들 혹은 비얼굴영들을 부정부류(negative class)로 간주하여 $l_i=-1$로 표시한다. 반면에 얼굴검출이 올바르게 된 얼굴영상들을 긍정부류(positive class)로 간주하고 $l_i=1$로 표시한다. SVM 분류기 모델은 다음과 같다.

$$f(x)=sgn(\sum_{i=1}^{L}\alpha_i l_i K(x_i,x)+b) \quad (1)$$

여기서 $K(\cdot)$는 입력 벡터를 특징공간에 삽입되는 커널 함수(kernel function)이며, α_i는 이중 최적화(dual optimization) 문제를 풀기 위한 라그랑지 곱수(Lagrange multipliers), sgn는 부호함수[22] 이다. 식

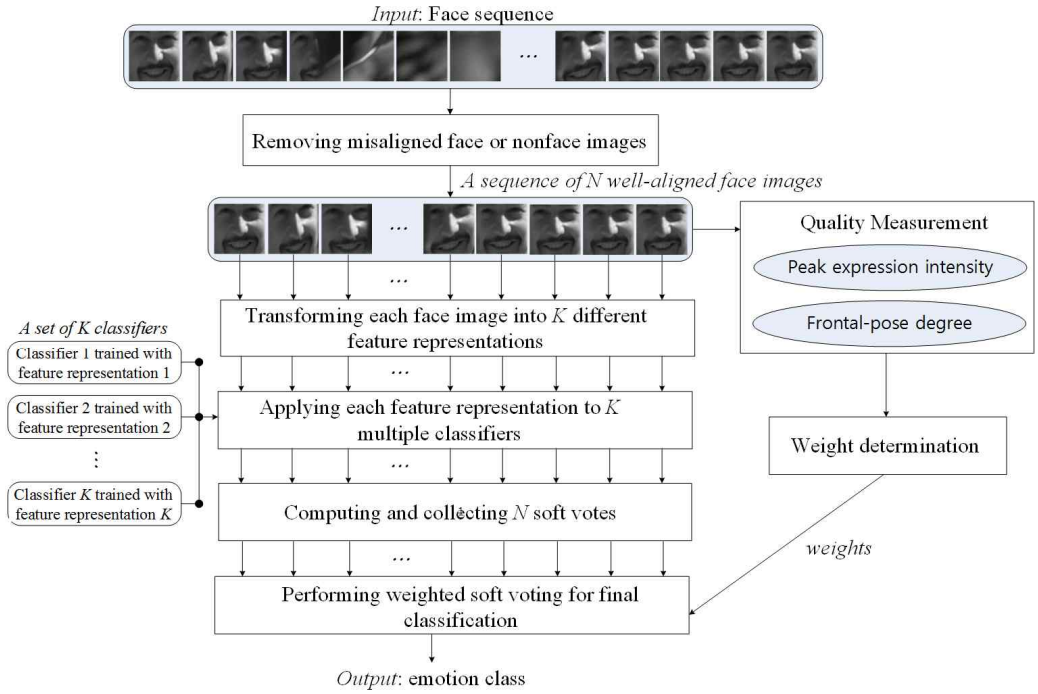

Fig. 1. Overall process of the proposed facial expression based emotion recognition framework for a given face sequence. Note that a face sequence is assumed to be obtained by applying face tracking and detection to a given video sequence.

(1)에서 표현된 SVM 분류기 모델 형성을 위해 우리는 커널 함수로 방사형 기저함수(radial basis function, RBF[24])를 사용했다.

식 (1)에 SVM 분류기를 형성하기 위해 1,000개의 긍정 얼굴 샘플들과 1,000개의 부정 얼굴 샘플 영상들을 수집하여 사용하였다. Fig. 2는 긍정 샘플들과 부정 샘플들로부터 구한 SVM 분류기 출력 값들을 [-1,1]의 범위로 정규화[23]한 확률 분포를 보여준다. Fig. 2에서와 같이 입력된 얼굴영상 x에 대해 SVM 분류기 출력 $f(x)$가 음수 값(negative value)을 갖는다면, x는 오정렬된 얼굴영상 혹은 비얼굴영상으로 간주되어 최종 얼굴 시퀀스에서 제거된다.

4. 가중 소프트 투표 분류 알고리즘

$\{\mathbf{I}_n\}_{n=1}^{N}$를 비디오 시퀀스에서 오정렬된 얼굴영상들 및 비얼굴영상들을 제거한 후 남은 N개의 얼굴영상들로 구성된 얼굴 시퀀스라 표기하자. 감정부류 라벨(emotion class label) ℓ_i는 i-번째 감정부류(예: 행복)를 나타낸다고 가정하자. 얼굴 시퀀스의 n번째 얼굴영상 $\mathbf{I}_n$에 대해 총 K개의 다른 특징표현들(feature representations)을 얻기 위해 참고문헌 [24]에서 기술한 특징추출(feature extraction) 방법들을 적용하였다. 여기서 f_m을 $\mathbf{I}_n$에서 계산된 m번째 특징표현이라 가정하자. 본 논문에서는 f_m을 얻기 위해 Local Binary Pattern(LBP)[25]와 Gabor Wavelet [26]을 활용한다. LBP 특징벡터 추출 연산 시 인접 픽셀의 개수 P와 원의 반경 R을 다음과 같이 (8,1), (8,2), (8,3), (16,1), (16,2)으로 조합하여 총 다섯 개의 LBP 특징표현들을 활용하였다. Gabor Wavelet에 경우 두 개의 스케일(scale)과 다섯 개의 방향(orientation)을 갖는 Gabor 필터들을 구성하여 총 열 개의 Gabor Wavelet 특징표현들을 활용하였다.

$h_m(\cdot)$을 m번째 특징표현 f_m으로 훈련된 분류기(classifier)라 하자, 여기서 $m=1,...,M$이고 M은 사용가능한 특징표현 유형(type)의 총 개수이다. 얼굴영상 $\mathbf{I}_n$이 $h_m(\cdot)$에 입력으로 주어질 때 해당 출력은 C차원을 가지는 벡터 $\left[h_m^{\ell_1}(\mathbf{I}_n), h_m^{\ell_2}(\mathbf{I}_n), ..., h_m^{\ell_C}(\mathbf{I}_n)\right]^{\mathrm{T}}$으로 표현된다. 여기서 C는 총 감정부류들의 개수이고 T

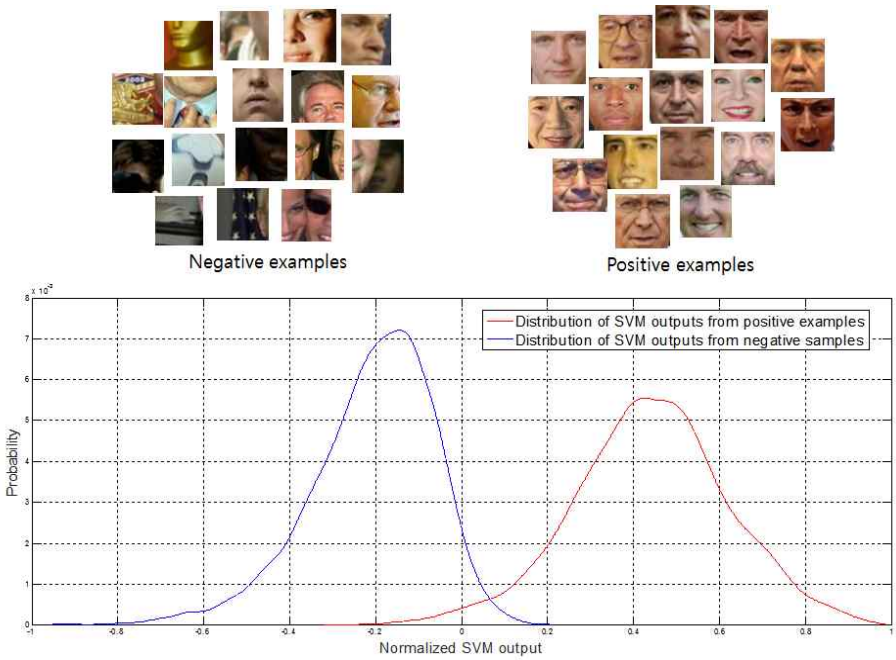

Fig. 2. Distribution of SVM classifier outputs trained with 1,000 positive and 1,000 negative face examples for filtering out misaligned- and non-face images.

는 행렬 및 벡터연산에서 전치 연산자(transpose operator)이다. 어떤 분류기 모델을 활용해도 제안 방법이 적용될 수 있도록, 각 분류기의 출력 범위는 $h_m^{\ell_i}(\mathbf{I}_n)\in[0,1]$이다. 또한 분류기 출력 $h_m^{\ell_i}(\mathbf{I}_n)$은 C개의 가능한 감정부류 라벨들 $\{\ell_1,...,\ell_C\}$에 대한 사후확률(posterior probability)의 추정치 $P(\ell_i|\mathbf{I}_n)$를 나타낸다. 얼굴영상 $\mathbf{I}_n$에 대한 소프트 투표 값은 다음과 같이 계산한다.

$$h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)=T\left(h_1^{\ell_i}(\mathbf{I}_n),h_2^{\ell_i}(\mathbf{I}_n),...,h_k^{\ell_i}(\mathbf{I}_n)\right)\text{for } i=1,...,C \quad (2)$$

여기서 $T(\cdot)$는 K개의 분류기 출력 값들을 결합하기 위한 결합규칙(combination rule)을 실행하는 함수이다. 예를 들어, $T(\cdot)$함수가 최대규칙(max rule)[27] 기반 결합을 수행할 경우 [즉, $h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)=\max_{m=1}^{K}h_m^{\ell_i}(\mathbf{I}_n)$], K개의 출력 값들 중 가장 큰 값을 가지는 분류기의 출력 값이 $h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)$이 된다. 본 논문에서는 $T(\cdot)$함수로서 평균규칙(average rule)[27], 최소규칙(min rule)[27], 최대규칙(max rule)[27]을 적용하여 평가한 결과 최대규칙 방식이 가장 우수한 성능을 보였다. 따라서 특별한 언급이 없는 한 이후 내용에서 $T(\cdot)$함수를 최대규칙을 수행하는 함수로 간주한다.

소프트 투표 값 $h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)$을 활용하여 가중치 소프트 투표 분류를 다음과 같이 실행한다.

$$\ell^*=\arg\max_i\sum_{n=1}^{N}w_n h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)\quad \text{for}\quad i=1,.....,C$$
$$\text{종속조건}\ \sum_{n=1}^{N}w_n=1\ \text{그리고}\ 0\le w_n\le 1 \quad (3)$$

여기서, w_n은 n번째 얼굴영상의 소프트 투표 값 $h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)$에 부여된 가중치(weight)이며, ℓ^*은 가중치 기반 소프트 투표 결합 $\sum_{n=1}^{N}w_n h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)$을 최대화하는 감정부류이다. 식 (2)에서 제시된 방법으로 계산된 소프트 투표 값과 해당 가중치 값을 결합하여 감정유형을 분류하는 방법을 본 논문에서 제안하는 '가중 소프트 투표 분류 알고리즘'이라 정의한다. 제안 방법에서는 감정인식 정확도 성능을 최적화하기 위해 더욱 신뢰성 있는 소프트 투표 $h_{\text{combined}}^{\ell_i}(\mathbf{I}_n)$들에 대해 더 큰 가중치를 부여하는 방법을 개발하였다. 이를 위해, 가중치값 w_n을 효과적으로 결정할 수 있는 알고리즘을 개발하였다. 가중치 결정 알고리즘에

대해서 다음 4.1절에서 설명한다.

4.1. 얼굴영상 품질 측정 활용 가중치 결정 알고리즘

식 (3)에서 제안한 가중치 기반 투표 분류 방법의 성능을 최적화하기 위해 가중치 w_n을 효과적으로 결정하는 방법을 제안한다. 제안 방법에서 얼굴영상의 품질은 '정면자세 정도(frontal pose degree)'와 '피크표정 강도(peak expression intensity)[28]' 요소들을 기준으로 측정한다. 여기서 '피크표정 강도'란 특정 감정(예: 분노, 행복)을 표현함에 있어 주어진 얼굴영상이 얼마만큼의 특성들을 표출하고 있는지 정도를 나타낸다(Fig. 3 예시 참조). 위에서 언급한 요소들을 측정하는 목적은 얼굴 시퀀스 내에서 정면자세 및 높은 피크표정 상태들을 동시에 가지는 얼굴 영상들의 소트프 투표 $h^{\ell_i}_{\mathrm{combined}}(\mathbf{I}_n)$에 더 큰 가중치 값들을 부여하기 위함이다. 앞서 언급한 두 가지 요소들은 다음과 같은 방식으로 측정 된다.

피크표현 강도(Peak expression intensity) : 실질적인 비디오 기반 감정인식 응용에서 얼굴영상들이 특정 감정을 대변하는 정도(degree)가 다르게 나타날 수 있다[28]. 따라서, 다중개의 얼굴영상들로 구성된 얼굴 시퀀스를 활용하여 표정인식을 수행하는 경우, 높은 표현정도를 가지는 비디오 프레임에서 얻어진 인식 결과가 최종 인식결과 결정에서 큰 가중치(영향력)을 갖도록 하는 것이 필요하다. 참고문헌 [28]에 내용을 기반으로 제안 방법에서 '피크표현의 강도'는 얼굴 영상에서 표현 된 감정상태(emotion state)가 무표정(neutral expression) 감정상태와 다른 정도로 정의된다. 이를 위해 본 논문에서는 고유공간(Eigenspace[29]) 기법을 기반으로 하는 '무표정 매니폴드 모델(Neutral Expression Manifold Model)'을 개발하였다. $\mathbf{T}^{\mathrm{neu}}$을 무표정을 갖는 얼굴영상들로 구성된 훈련집합(training set)이라 가정하자. 본 논문에서는 공식 표정인식 데이터베이스들인 FER-2013[30]과 Multi-PIE[31]에서 무표정 얼굴영상들 5,120개를 수집하여 형성하였다. 참고문헌 [29]에서 제시한 방법에 따라 $\mathbf{T}^{\mathrm{neu}}$를 일반적인 얼굴 고유공간을 형성하기 위한 훈련집합으로 활용하면 무표정 상태의 얼굴특성들의 주고유벡터(principal eigenvectors)들로 구성된 기저집합(spanning set)으로 구성되는 매니폴드(manifold) $\boldsymbol{\Phi}^{\mathrm{neu}}$를 얻을 수 있다. 얼굴 시퀀스 집합 $\{\mathbf{I}_n\}_{n=1}^{N}$ 에서 각 얼굴영상 $\mathbf{I}_n$의 피크표현 강도 Q_n^{peak}는 다음과 같이 계산된다.

$$Q_n^{\mathrm{peak}} = \frac{R(\mathbf{I}_n, \Phi^{\mathrm{neu}}) - \min\left(\left\{R(\mathbf{I}_n, \Phi^{\mathrm{neu}})\right\}_{n=1}^{N}\right)}{\max\left(\left\{R(\mathbf{I}_n, \Phi^{\mathrm{neu}})\right\}_{n=1}^{N}\right) - \min\left(\left\{R(\mathbf{I}_n, \Phi^{\mathrm{neu}})\right\}_{n=1}^{N}\right)} \tag{4}$$

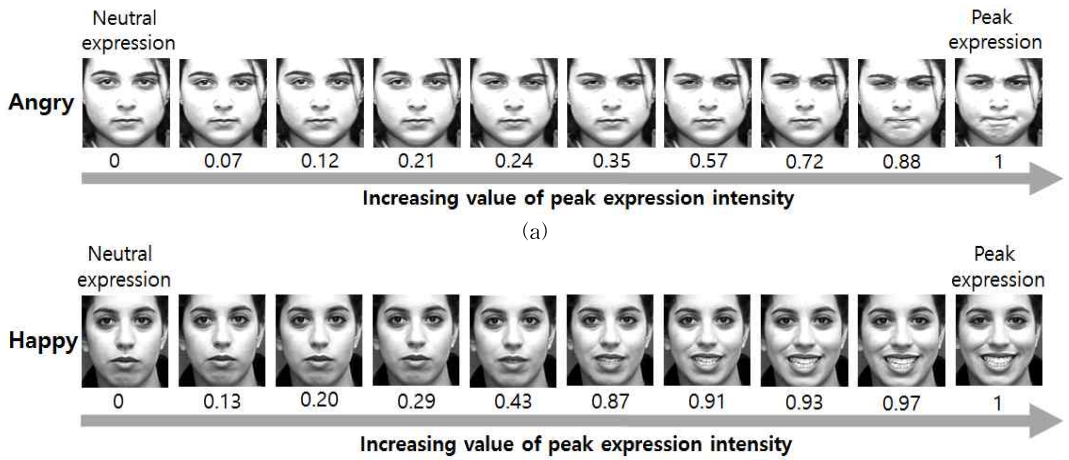

Fig. 3. Visualized illustration for measuring the "peak expression intensity" for each facial image contained in a face sequence. Facial images within a face sequence are sorted in decreasing order of corresponding "peak expression intensity". Note that the value below each facial image represents the value for the peak expression intensity measured via using Eq. (4) and (5). (a) Face sequence with angry emotion class and (b) face sequence with happy emotion class.

여기서 $R(\cdot)$은 얼굴영상 $\mathbf{I}_n$에 대해 무표정 감정상태와의 비유사성(dissimilarity)을 측정하는 재구성 에러(reconstruction error)이며 다음과 같이 계산된다.

$$R(\mathbf{I}_n, \Phi^{\text{neu}}) = \| \mathrm{v}_n - (\Phi^{\text{neu}})\mathrm{e}_n \|_2^2 \tag{5}$$

그리고 $\mathrm{e}_n = (\Phi^{\text{neu}})^{\mathrm{T}} \mathrm{v}_n$

여기서 v_n은 2차원 영상 $\mathbf{I}_n$의 1차원 열 벡터(column vector) 표현이고(영상 화소 값들을 행(혹은 열) 방향으로 스캐닝(scanning)해서 형성) e_n은 주성분들을(principal components) 원소로 가지는 벡터이며 $\| \cdot \|_2$은 L_2 놈(norm)을 나타낸다. 식 (4)에서 Φ^{neu}의 값의 범위는 [0, 1]이며 Fig. 3에 예시와 같이 '1'의 가까운 값을 가질수록 특정 감정(분노 혹은 행복)을 대변하는 표정상태가 뚜렷해짐을 볼 수 있다.

정면자세 정도 (Frontal-pose degree) : 얼굴영상 $\mathbf{I}_n$의 정면자세 정도를 측정하기 위해 정면자세 부공간(frontal-pose subspace)[33] 모델을 형성하였다. 이를 위해 CMU-PIE[34]와 FERET[35] 얼굴인식 데이터베이스로부터 수집된 정면 자세 얼굴영상들을 활용하였다. 정면 자세를 갖는 총 1,530개의 얼굴영상(101명의 인식 대상자)들이 정면자세 부공간 형성을 위한 주성분분석(principal component analysis, PCA)을 실행하는데 활용되었다. 형성된 정면자세 부공간을 Φ^{frt}로 나타낼 때, Φ^{frt}는 부공간 평균(subspace mean)과 고유벡터들(eigenvectors)로 구성된다. 얼굴영상 $\mathbf{I}_n$의 정면자세 정도는 다음과 같이 계산된다.

$$Q_n^{\text{frt}} = 1 - \frac{R(\mathbf{I}_n, \Phi^{\text{frt}}) - \min\left(\left\{R(\mathbf{I}_n, \Phi^{\text{frt}})\right\}_{n=1}^{N}\right)}{\max\left(\left\{R(\mathbf{I}_n, \Phi^{\text{frt}})\right\}_{n=1}^{N}\right) - \min\left(\left\{R(\mathbf{I}_n, \Phi^{\text{frt}})\right\}_{n=1}^{N}\right)} \tag{6}$$

여기서 $R(\bullet)$은 $\mathbf{I}_n$이 정면자세 부공간을 구성하는 고유벡터들과 고유값들을 이용해 선형조합(linear combination)으로 재구성될 때 재구성 오차를 계산하는 함수이며 식 (5)에서 제시한 방법으로 수행된다. Q_n^{frt}는 [0, 1]의 값을 가지며 1에 가까울수록 정면자세를 갖는 얼굴영상으로 판단된다.

식 (4)와 (6)의 Q_n^{peak}와 Q_n^{frt}를 활용해 아래와 같이 $\mathbf{I}_n$의 가중치를 계산한다.

$$w_n = \frac{\alpha Q_n^{\text{peak}} + (1-\alpha) Q_n^{\text{frt}}}{\sum_{n=1}^{N} \alpha Q_n^{\text{peak}} + (1-\alpha) Q_n^{\text{frt}}} \tag{7}$$

여기서 α는 피크표정 강도 Q_n^{peak}과 정면자세 정도 Q_n^{frt}의 절충관계(trade-off)를 조정하는 매개변수(parameter)이며 본 논문에서는 $\alpha = 0.65$로 설정하여 피크표정 강도를 좀 더 부각하였다. 식 (3)에서 제안한 가중치 소프트 과반수 투표 분류 방법의 종속조건 $\sum_{n=1}^{N} w_n = 1$과 $0 \le w_n \le 1$을 만족하기 위해 식 (7)에 w_n을 다음과 같이 정규화하여 최종적으로 결정한다.

$$w_n = \frac{w_n}{\sum_{n=1}^{N} w_n} \tag{8}$$

5. 실험환경 및 결과

제안한 가중치 소프트 투표 분류 알고리즘을 검증하기 위해 공개된 표정인식 데이터베이스 CK+ DB [18]에 대해 실험을 수행했다. 실험에 사용된 영상은 intra-face SW [36]의 안면표식검출 기술(facial landmark detection method)을 사용하여 두 눈의 위치를 기준으로 영상에서 얼굴을 포함하는 영역만을 분리한 후 정렬하였다 [37]. Fig. 4는 얼굴표식검출 결과 얼굴영상들의 예시들을 보여준다.

CK+ DB는 총 123명의 인식 대상자들로 이루어져 있으며, 593개의 비디오 시퀀스들을 포함한다. 593개의 비디오 시퀀스들에서 118명의 인식 대상자로 구성된 325개의 비디오 시퀀스들을 일곱 개의 감정부류들에 따라 선택하였다[18]. 선택된 비디오 시퀀스들에 Fig. 4와 같이 안면표식검출 기술을 적용하여 얼굴 시퀀스들을 생성하였고, 분노(anger), 경멸(contempt), 혐오감(disgust), 두려움(fear), 행복(happy), 슬픔(sadness), 놀람(surprise) 감정부류에 대해 각각 45, 18, 58, 25, 69, 28, 82개의 얼굴 시퀀스들로 구성하였다. Fig. 5는 실험에 사용된 얼굴 시퀀스들 중 일부 예시들을 제시한다. 각 얼굴 시퀀스 당 한 개의 피크표정(peak expression)을 갖는 얼굴영상이 존재한다. 신뢰성 있는 실험결과를 얻기 위해 [38, 39]에서 제시한 방법과 동일하게 Leave-one-subject-out(LOSO) 교차검증(cross validation)을 평가 프로토콜(evaluation protocol)로 활용 하였다. LOSO

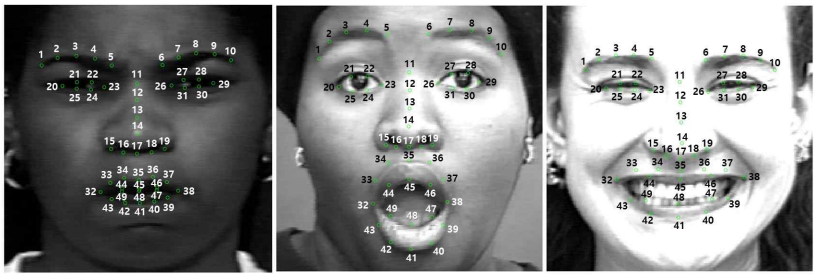

Fig. 4. Facial landmark points detected using the method in [36]. For each facial iamge, the coordinates of the left eye and the right eye are obtained by averaging the coordinates of the facial landmarks No. 20–25 and the facial landmarks No. 26–31, respectively.

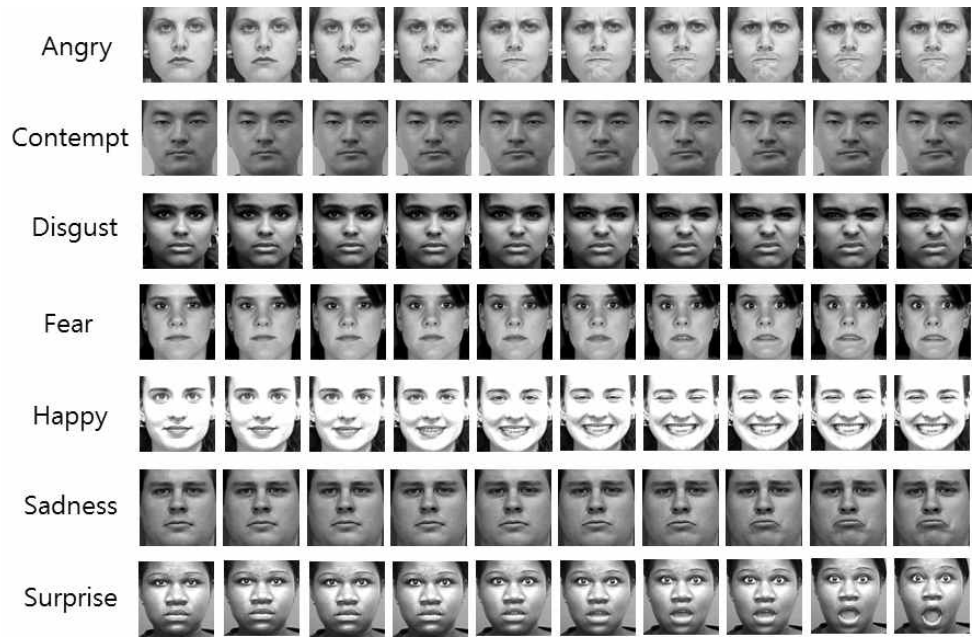

Fig. 5. Examples of face sequences, each corresponding to one of the seven emotion classes, used in our experimentations.

에서는 118명의 인식 대상자 중 117명의 얼굴 시퀀스들이 훈련 집합(training set)으로 활용되고, 나머지 얼굴 시퀀스들은 테스트(test) 과정에 사용되었다. 이 과정을 118명의 인식 대상자 모두에게 반복하여 반복 횟수에 대한 평균성능을 표정인식 성능으로 도출하였다.

제안 방법에서 $h_m(\cdot)$을 형성하기 위해 지원벡터기계(SVM) 분류기 모델을 활용하였다. 표정인식을 위해 전통적으로 많이 활용되고 있는 텍스처 특징(texture feature) 기반 방법들이 제안한 방법과 비교를 위해 적용되었다. 이를 위해 LBP(Local Binary Pattern), 가보 웨이블렛(Gabor wavelet) 및 LPQ (Local Phase Quantization) 방법들을 비교하고 평가하였다. [40]에서 사용된 LBP 특징들을 추출하기 위해, uniform LBP 연산(operation) 실행 시 매개변수 $P = 8$, $R = 1$로 설정하여 사용하였다. LPQ 특징을 추출하기 위해 5×5 neighborhoods를 사용하였다 [41]. 가보 웨이블렛 특징을 추출하기 위해 다섯 개의 크기와 여덟 개의 방향을 사용하여 가보필터뱅크(Gabor filter bank) 집합을 구성하였다[44]. LBP와 LPQ 특징을 추출하기 위해 [40,43]에서와 같이 얼굴 영상들은 64×64를 갖는 영상 크기로 재조정 하였다.

Table 1. Comparison of FER recognition rates on CK+ DB with texture feature based video FER methods

FER method	Recognition rate (%)
LBP[40]	79.75
LPQ[43]	82.91
Gabor[44]	82.20
Proposed method	91.20

CK+ DB에 대한 비교 실험결과들이 Table 1에 제시된다. LBP, LPQ 및 Gabor Wavelet 기반 비디오 표정인식 방법들은 79.75%에서 82.91%로 비교적 낮은 성능을 보였다. 반면, 제안한 방법의 경우에는 약 91.20%의 가장 높은 표정인식 성능을 달성하였다.

제안 방법의 우수성을 추가적으로 검증하기 위해 최근에 발표된 표정인식 방법들[44,45,46,47,39]과 비교하는 실험을 수행하였다. Table 2에서 볼 수 있듯이 참고문헌 [39]의 표정인식 방법이 인식 성능 92.3%로 가장 높은 표정인식 정확도를 보였고, 제안하는 방법은 91.2%에 성능으로 두 번째로 높은 표정인식 정확도를 보였다. 하지만 여기서 주목할 점은 참고문헌 [39]의 표정인식 방법은 수작업(manual operation) 방법으로 얼굴영상들을 정렬(alignment)하였고 심지어 피크표정(peak expression)을 갖는 얼굴영상들을 선택하는 과정에서도 반자동(semi-automatic) 기법을 적용하였다. 따라서 [39]에서 보고된 표정인식 성능 결과는 실질적인 감정인식 응용에서의 표정인식 성능을 신뢰성 있게 대변한다고 생각하기 어렵다. 반면에 제안 방법이 달성한 표정인식 성능은 얼굴검출 및 정렬, 소프트 투표 가중치 결정, 감정 분류 등 제안방법을 실행하기 위한 모든 과정들을 자동으로 수행하여 획득한 성능이기 때문에 실제 응용에 적용한다는 측면에서 가장 우수한 표정인식 정확도를 달성했다고 볼 수 있다. 실행속도 측면에서 제안방법의 우수성을 검증하기 위해 컴퓨터 사양은 window 10의 운영체제, CPU는 3.40GHz I7-6700, RAM은 16GB, 그래픽카드는 NVIDIA GeForce GTX 950을 활용하여 수행시간을 측정하였다. 제안방법은 얼굴영상 당 0.026초의 수행시간이 필요한 것으로 판명되었고 이는 최신 방법들과 비교했을 때 상대적으로 빠른 고속처리가 가능함을 나타낸다.

6. 결 론

본 논문에서는 감정인식을 위한 비디오 시퀀스 기반 가중치 소프트 투표 분류 알고리즘과 가중치를 최적화하기 위한 가중치 결정 알고리즘을 제안하였다. 오정렬되거나(misaligned) 비얼굴영상(non-face image)이 제거된 얼굴 시퀀스는 상이한 특징표현들로 학습된 다중 분류기들에 각각 입력되어 얼굴 시퀀스에 포함된 각 얼굴영상의 특정 감정유형(emotion class)에 대한 소프트 투표(soft voting)값들을 얻은 뒤, 얼굴 시퀀스에 포함된 얼굴영상들의 품질 척도인 '표정의 피크표현 강도(peak expression intensity)'

Table 2. Comparisons with the recent advances in FER on CK+ DB

Method	Cross validation	No. subjects	No. expressions	Recognition Rate (%)	Computation time (sec)
Proposed	LOSO[1]	118	7[4]	91.2	0.026[7]
LDN_K[44]	10-fold[2]	118	7[4]	82.3	0.037[7]
LDN_G[44]	10-fold[2]	118	7[4]	89.3	0.037[7]
Intra-class variation reduction[47]	LOSO[1]	118	7[4]	90.5	0.15[7]
Intra-class variation reduction[47]	10-fold[2]	118	7[4]	89.6	0.15[7]
STLMBP-C[39]	LOSO[1]	118	7[4]	92.3	Not stated
Component-based[46]	LOSO[1]	106	7[5]	89.2	Not stated
Neutral-independent Geometric features[45]	LOSAO[3]	Not stated	7[5]	82.3	Not stated
Neutral-independent Geometric features[45]	LOSAO[3]	Not stated	7[6]	73.4	Not stated

[1]leave-one-subject-out cross validation, [2]10-fold subject-independent cross validation
[3]denotes leave-one-sample-out cross validation, [4]six basic expressions + contempt
[5]six basic expressions (contempt is excluded), [6]six basic expressions + neutral (contempt is excluded)
[7]average computation time per facial image

와 '정면자세 정도(frontal pose degree)'를 정량화하여 구한 가중치와 해당 소프트 투표 값들을 결합하여 최종 감정부류를 결정하였다. 최신 비디오 기반 표정인식 기술들과의 비교실험 결과 제안방법은 91.2%의 높은 감정인식 성능을 달성하였다.

REFERENCE

[1] P. Michel and R. El Kaliouby, "Real Time Facial Expression Recognition in Video Using Support Vector Machines," *Proceedings of the 5th International Conference on Multimodal Interfaces (Association for Computing Machinery)*, pp. 258-264, 2003.

[2] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and Dynamic 3D Facial Expression Recognition: A Comprehensive Survey," *Image and Vision Computing*, Vol. 30, No. 10, pp. 683-697, 2012.

[3] S. Taheri, Q. Qiu, and R. Chellappa, "Structure-Preserving Sparse Decomposition for Facial Expression Analysis," *IEEE Transactions on Image Processing*, Vol. 23, No. 8, pp. 3590-3603, 2014.

[4] Y.L. Tian, T. Kanade, and J.F. Cohn, *Facial Expression Analysis, Handbook of Face Recognition*, Springer, New York, pp. 247-276, 2005.

[5] R.V. Yampolskiy, *Artificial Superintelligence: a Futuristic Approach*, Chemical Rubber Company Press, London, 2015.

[6] Microsoft Azure, https://www.projectoxford.ai/demo/Emotion#detection (accessed July, 10, 2017).

[7] Affectiva, https://www.affectiva.com (accessed July, 10, 2017).

[8] J. Stallkamp, H.K. Ekenel, and R. Stiefelhagen, "Video-based Face Recognition on Real-World Data on Real-world Dataset," *Proceeding of IEEE International Conference on Computer Vision*, pp. 1-8, 2007.

[9] Y. Zhang and A.M. Martinez, "A Weighted Probabilistic Approach to Face Recognition from Multiple Images and Video Sequences," *Image Vision Computing*, Vol. 24, No. 6, pp. 626-638, 2006.

[10] A. Hamid and M. Pietikainen, "From still Image to Video-based Face Recognition: An Experimental Analysis," *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 813-818, 2004.

[11] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous Facial Feature Tracking and Facial Expression Recognition," *IEEE Transactions on Image Processing*, Vol. 22, No. 7, pp. 2559-2573, 2013.

[12] J.J. Lien, T. Kanade, J.F. Cohn, and C. Li, "Detection, Tracking, and Classification of Action Units in Facial Expression," *Journal of Robotics and Autonomous Systems*, Vol. 31, No. 3, pp. 131-146, 2000.

[13] Y. Chang, C. Hu, R. Feris, and M Turk, "Manifold Based Analysis of Facial Expression," *Image Vision Computing*, Vol. 24, No. 10, pp. 605-614, 2006.

[14] G. Zhao and M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 29, No. 6, pp. 915-928, 2007.

[15] B. Jiang, F. Valstar, and M. Pantic, "Action Unit Detection Using Sparse Appearance Descriptors in Space-Time Video Volumes," *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 314-321, 2011.

[16] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, Vol. 44, No. 2, pp. 161-174, 2014.

[17] D. Ruta and B. Gabrys, "Classifier Selection for Majority Voting," *Information Fusion,* Vol. 6, No. 1, pp. 63-81, 2005.

[18] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih,

and Z. Ambadar, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-specified Expression," *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 94-101, 2010.

[19] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceeding of IEEE International Conference on Computer Vision Pattern Recognition*, pp. I-511-I-518, 2001.

[20] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," Carnegie Mellon University, Pittsburgh, PA, Technical Report CMU-CS-91-132, 1991.

[21] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 568-573, 2005.

[22] J.A. Davis, D.E. McNamara, D.M. Cottrell, and J. Campos, "Image Processing with the Radial Hilbert Transform: Theory and Experiments," *Optics Letters*, Vol. 25, No. 2, pp. 99-101, 2000.

[23] A. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Biometric Systems," *Journal of Pattern Recognition*, Vol. 38, No. 12, pp. 2270-2285, 2005.

[24] J.Y. Choi, Y.M. Ro, and K.N. Plataniotis, "Color Local Texture Features for Color Face Recognition," *IEEE Transactions on Image Processing*, Vol. 21, No. 3, pp. 1366-1380, 2012.

[25] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Pattern: Application to Face Recognition," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 28, No. 12, pp. 2037-2041, 2006.

[26] C. Liu and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Transactions on Image Processing*, Vol. 11, No. 4, pp. 467-476, 2002.

[27] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 20, No. 3, pp. 226-239, 1998.

[28] M. Suk and P. Balakrishnan, "Real-time Mobile Facial Expression Recognition System-A Case Study," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 132-137, 2014.

[29] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 24, No. 6, pp. 780-788, 2002.

[30] P.L. Carrier, A. Courville, I.J. Goodfellow, M. Mirza, and Y. Bengio, *FER-2013 Face Database*, Technical Report, 2013.

[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, Vol. 28, No. 5, pp. 807-813, 2010.

[32] P. Michel and R.E.l. Kaliouby, "Real Time Facial Expression Recognition in Video Using Support Vector Machines," *Proceedings of the 5th Association for Computing Machinery International Conference on Multimodal Interfaces,* pp. 258-264, 2003.

[33] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 24, No. 6, pp. 780-788, 2002.

[34] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 25, No. 12, pp. 1615-1618, 2003.

[35] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms," *IEEE*

Transactions on Pattern Analysis Machine Intelligence, Vol. 22, No. 10, pp. 1090-1104, 2000.

[36] Intra-Face, http://humansensing.cs.cmu. edu/intraface (accessed June, 23, 2017).

[37] Y. Tian, "Evaluation of Face Resolution for Expression Analysis," *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, pp. 82-88, 2004.

[38] H.W. Kang, K.T. Lim, and C.H. Won, "Learning Directional LBP Features and Discriminative Feature Regions for Facial Expression Recognition," *Journal of Korea Multimedia Society*, Vol. 20, No. 5, pp. 748-757, 2017.

[39] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Spatiotemporal Local Monogenic Binary Patterns for Facial Expression Recognition," *IEEE Signal Processing Letters*, Vol. 19, No. 5, pp. 243-246, 2012.

[40] M. Huang, Z. Wang, and Z. Ying, "A New Method for Facial Expression Recognition Based on Sparse Representation Plus LBP," *IEEE International Congress on Image and Signal Processing*, pp. 1750-1754, 2010.

[41] W. Zhen and Y. Zilu, "Facial Expression Recognition Based on Local Phase Quantization and Sparse Representation," *Proceeding of IEEE International Conference on Natural Computation*, pp. 222-225, 2012.

[42] S. Zafeiriou and M. Petrou, "Sparse Representation for Facial Expression Recognition via l1 Optimization," *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 32-39, 2010.

[43] M.F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The First Expression Recognition and Analysis Challenge," *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 921-926, 2011.

[44] A.R. Rivera, J.R. Castillo, and O. Chae, "Local Directional Number Pattern for Face Analysis," *IEEE Transactions on Image Processing*, Vol. 22, No. 5, pp. 1740-1752, 2013.

[45] A. Saeed, A. Al-Hamadi, and R. Niese, "Neutral-independent Geometric Features for Facial Expression Recognition," *Proceeding of IEEE International Conference on Intelligent Systems Design and Applications*, pp. 842-846, 2012.

[46] S. Taheri, V.M. Patel, and R. Chellappa, "Component-based Recognition of Faces and Facial Expressions," *IEEE Transactions on Affective Computing*, Vol. 23, No. 8, pp. 360-371, 2013.

[47] S.H. Lee, K.N. Plataniotis, and Y.M. Ro, "Intra-Class Variation Reduction Using Training Expression Images for Sparse Representation Based Facial Expression Recognition," *IEEE Transactions on Affective Computing*, Vol. 5, No. 3, pp. 340-351, 2014.

김 경 태

2016년 중원대학교 학사
2017년 현재 한국외국어대학교 컴퓨터.전자공학부 석사
관심분야 : 머신러닝, 패턴인식, 영상처리

최 재 영

2011년 KAIST 전기및전자공학과 박사
2008년~2009년 토론토대학 연구원
2011년~2012년 토론토대학 연구원
2012년~2013년 펜실베니아대학 연구원
2013년~2014년 삼성전자 책임연구원
2014년~2016년 중원대학교 의료공학과 조교수
2016년~현재 한국외국어대학교 컴퓨터.전자공학부 조교수
관심분야 : 딥 러닝, 머신러닝, 패턴인식, 영상처리