Multi-Objective based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition

Article *in* IEEE Transactions on Affective Computing · April 2017

DOI: 10.1109/TAFFC.2017.2695999

Wissam Baddar
Samsung Advanced Institute of Technology

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2695999, IEEE Transactions on Affective Computing

Multi-Objective based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition

Dae Hoe Kim*, Wissam J. Baddar*, Jinhyeok Jang and Yong Man Ro, *Senior Member, IEEE*

Abstract— Facial expression recognition (FER) is increasingly gaining importance in various emerging affective computing applications. In practice, achieving accurate FER is challenging due to the large amount of inter-personal variations such as expression intensity variations. In this paper, we propose a new spatio-temporal feature representation learning for FER that is robust to expression intensity variations. The proposed method utilizes representative expression-states (e.g., onset, apex and offset of expressions) which can be specified in facial sequences regardless of the expression intensity. The characteristics of facial expressions are encoded in two parts in this paper. As the first part, spatial image characteristics of the representative expression-state frames are learned via a convolutional neural network. Five objective terms are proposed to improve the expression class separability of the spatial feature representation. In the second part, temporal characteristics of the spatial feature representation in the first part are learned with a long short-term memory of the facial expression. Comprehensive experiments have been conducted on a deliberate expression dataset (MMI) and a spontaneous micro-expression dataset (CASME II). Experimental results showed that the proposed method achieved higher recognition rates in both datasets compared to the state-of-the-art methods.

Index Terms— Facial expression recognition (FER), expression intensity variation, spatio-temporal feature representation, deep learning, long short-term memory (LSTM)

1 Introduction

Facial expression is considered important for non-verbal communication since it can convey the internal states, emotions and intentions of humans [1]. In recent years, researchers in computer vision, affective computing and human computer interaction fields have been trying to automatically recognize and interpret facial expressions [2], [3]. This is mainly attributed to the multitude of potential applications, such as interactive agents, fatigue measurement or even lie detection [4], [5].

While many research efforts have been devoted to explore facial expression recognition (FER), most of the previous researches have adopted hand-crafted features. Those methods can be broadly divided into two groups; static FER [6], [7], [8], [9], [10] and dynamic FER [11], [12], [13], [14], [15]. Static (frame-based) FER relied only on static facial features obtained by extracting hand-crafted features from the selected peak (apex) expression frames of video sequences. The effectiveness of various types of appearance features (e.g., local binary patterns (LBPs) [16], local phase quantization (LPQ) [17], Gabor wavelets [18], etc.) has been investigated. These hand-crafted features could cause a feature confusion that could occur between the facial expression and the facial identity [19]. To reduce the effect of subject identity on the FER performance, some methods adopted appearance difference [20], [21], geometrical difference [20], [22] or both differences [23] between the query face and the neutral face of a subject. However, in real-world applications, it is generally difficult to assume that the identity of subject is known and the neutral face image of subject is available beforehand [24], [25]. Moreover, the performance of such methods could be degraded when non-apex frames are used or when the expression intensity is small (i.e., subtle and micro-expressions).

Inspired by the fact that facial expressions are inherently a dynamic process [11], research efforts have utilized spatio-temporal features in order to capture the expression dynamics in facial expression sequences. Spatio-temporal features, such as local binary patterns from three orthogonal planes (LBP-TOP) [12] and local phase quantization from three orthogonal planes (LPQ-TOP) [11], [13] were obtained by extending LBP [16] and LPQ [17] into the three orthogonal planes (XY plane: appearance, XT planes: horizontal motion, and YT plane: vertical motion) [12], [13]. Although the spatio-temporal features have shown an improvement on the FER performance compared to the static (frame-based) features, they suffer from the following drawbacks. First, the spatial appearance features extracted from XY plane could encode the subject identity which is not desirable for FER [14]. Second, the extracted dynamic features could be negatively affected by the different characteristics of the facial ex-

- *The authors are with the School of Electrical Engineering, KAIST, Daejeon, Republic of Korea. E-mail: {dhkim10, wisam.baddar, jjh6297, ymro}@kaist.ac.kr.*

* *Both authors contributed equally to this manuscript.*

pression (e.g., different transition duration and/or transition type such as onset or offset [13]). Added to that, these spatio-temporal features adopt temporal normalization to obtain expression sequences with a fixed number of frames. The temporal normalization methods could result in the loss of temporal scale information.

The recent successes of deep learning in different fields of computer vision have prompted researchers to investigate in utilizing deep learning methods in FER [19], [26], [27], [28], [29], [30], [31], [32]. In [30], [31], local textural patterns, namely 'micro-action-patterns', were learned by a convolutional layer at apex expression frames. The learned micro-action-patterns were clustered and passed to a restricted Boltzmann machine to learn a spatial feature representation of the expression. In [26], a deep convolutional neural network (CNN) was used to learn a spatial feature representation. By relying on the spatial feature only, the method did not utilize facial expression dynamics were while performing the FER task, which can limit the performance at non-apex frames or frames of subtle expressions. In [19], a 3D CNN was utilized to learn spatio-temporal features from deformable facial action parts. In [32], a 3D CNN was also used in order to learn the spatio-temporal appearance features of the sequence. In [32], a temporal geometric feature was jointly learned in order to reduce the effect of the identity on the learned spatio-temporal appearance features. Although 3D CNNs can capture the dynamics of the expression, the 3D convolution is computationally expensive. Moreover, a large number of parameters (weights and biases) need to be learned, which makes the learning processes harder [33], [34]. To reduce complexity and the number of learned parameters, a small number of frames need to be used as an input for the 3D CNN. This could cause a loss of information on facial expression dynamics [35]. Subtle expression dynamics that occur on micro-expression sequences could be missed due to the short temporal range of the 3D CNN. In other methods [27], [28], [29], a recurrent neural network (RNN) was used with the conventional CNN to encode dynamics in the sequence for classification of facial expression [27] and prediction of the arousal and valence scores of emotions [28], [29]. The methods showed that the architectures of CNN with RNN can improve recognition performances compared to conventional CNN.

While many research efforts have been conducted for FER, achieving accurate and effective FER is still a challenging problem. This is mainly attributed to the large amount of variations in the imaging conditions and the inter-personal variations such as expression intensity variations [31]. In practice, facial expressions appear in a wide range of expression intensity, i.e., expression intensity can span from subtle micro-expression motions to exaggerated expressions [35], [36]. Therefore, a robust FER method to expression intensity variation is of paramount importance for practical FER. In this paper, we propose a new spatio-temporal feature representation learning for FER that is robust to expression intensity variations. The contributions of this paper are summarized as follows

1. New spatio-temporal feature representation learning method to robust expression intensity variations is proposed. Differing from the simple CNN and RNN architecture [27] which only utilized expression class information, the proposed method utilizes the representative expression-state information (e.g., onset, apex and offset of expressions) in the network training with proposed objective terms. It has been widely accepted that the expression-states is useful for facial analysis [35], [36]. By incorporating the representative expression-states, the proposed method mitigates the problem of the expression intensity variations and varying expression duration, and at the same time, focuses on improving expression class separability of the learned feature representation. The proposed learning method consists of two parts. In the first part, spatial feature representation of the facial expression is learned using the CNN with representative expression-states (i.e., onset, onset to apex transition, apex, apex to offset transition and offset). The CNN adopts the expression-states in the objective function to regulate the learning of the spatial feature representation. So, the expression class separability of the learned spatial feature representation is improved. In the second part, temporal characteristics of the spatial feature representation in the first part are learned with a long short-term memory (LSTM) [37]. The temporal sequence learning is devised to learn time scale dependent information in facial expression sequences that have varying number of frames.
2. To incorporate the representative expression-state information in the network training and improve the expression class separability of feature representation in feature space, five objective terms are devised. Two objective terms are devised to deal with expression-variations, namely minimizing expression classification error, and minimizing expression intra-class variations. Those objective terms focus on increasing the expression class separability of the learned spatial feature representation, and mitigating the effect of intra-class variations, such as subject identity variations. In addition, two objective terms are devised to deal with expression-state variations, namely minimizing expression-state classification error and minimizing expression-state variations. As a result, each expression class would contain distinct clusters of the expression-state feature representation, which can improve sequence learning of the expression. To preserve the expression-state continuity for further sequence learning, the fifth objective function is devised. The objective term for the expression-state continuity is useful to efficiently encode intermediate expression-states (e.g., it is useful when some of expression-states are missed).

To the best of our knowledge, this work is the first attempt to devise objective terms to utilize the expression state information in deep learning for facial expression

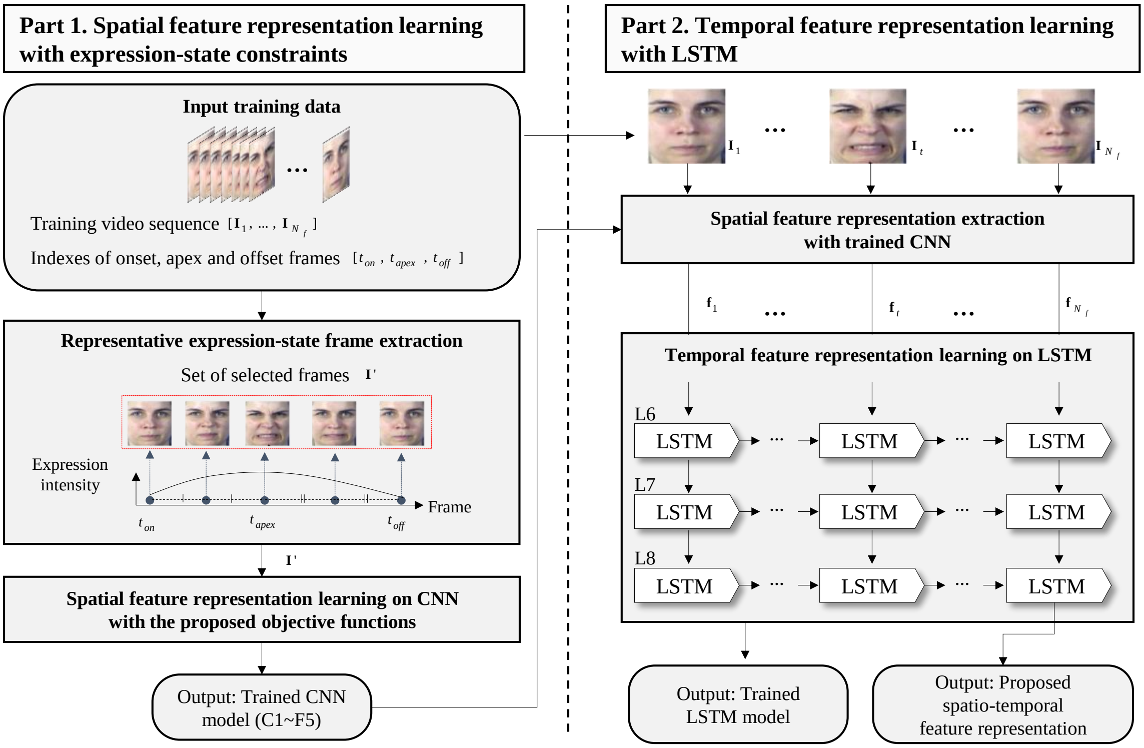

Part 1. Spatial feature representation learning with expression-state constraints
Input training data
Training video sequence $[\mathbf{I}_1, \dots, \mathbf{I}_{N_f}]$
Indexes of onset, apex and offset frames $[t_{on}, t_{apex}, t_{off}]$
Representative expression-state frame extraction
Set of selected frames $\mathbf{I}'$
Expression intensity
Frame
t_{on}
t_{apex}
t_{off}
$\mathbf{I}'$
Spatial feature representation learning on CNN with the proposed objective functions
Output: Trained CNN model (C1~F5)
Part 2. Temporal feature representation learning with LSTM
$\mathbf{I}_1$
$\mathbf{I}_t$
$\mathbf{I}_{N_f}$
Spatial feature representation extraction with trained CNN
$\mathbf{f}_1$
$\mathbf{f}_t$
$\mathbf{f}_{N_f}$
Temporal feature representation learning on LSTM
L6
L7
L8
LSTM
Output: Trained LSTM model
Output: Proposed spatio-temporal feature representation

Fig. 1. Overview of the proposed spatio-temporal feature representation learning for facial expression recognition robust to expression intensity variations.

recognition. Extensive and comprehensive experiments have been conducted on the MMI dataset [36] for deliberate expressions and the CASME II dataset [35] for micro-expressions. Experimental results from both datasets show that the expression class separability of the learned feature representation is improved in both types of expression intensity (i.e., deliberate expressions and micro-expressions). As a result, the proposed method outperforms state-of-the-art methods on both datasets in terms of the recognition rate. Moreover, an inter-dataset experiment shows that the FER with the learned feature representation maintains a comparable performance under intra-dataset conditions. These results indicate that the proposed method can learn versatile spatio-temporal feature representation for both deliberate expression and micro-expression.

The remainder of this paper is organized as follows. Section 2 details the proposed facial expression recognition with learned spatio-temporal feature representation. Section 3 presents experimental results to verify the effectiveness of the proposed spatio-temporal feature representation. Finally, conclusions are drawn in section 4.

2 FACIAL EXPRESSION RECOGNITION WITH LEARNED SPATIO-TEMPORAL FEATURE REPRESENTATION

2.1 Overview of the Proposed Method

Fig. 1 shows the overview of the proposed spatio-temporal feature representation learning for FER. The proposed learning for spatio-temporal feature representation consists of two parts: 1) spatial feature representation learning with expression-state constraints and 2) consecutive temporal-feature representation learning with LSTM. In the first part, representative frames of different expression-states (i.e., onset, onset to apex transition, apex, apex to offset transition and offset) of the facial expression sequence are utilized in the learning. From the ground truth of representative expression-state frames, spatial feature representations are learned by a CNN with an objective function with multiple objective terms, which are robust to expression intensity variations. The objective terms are devised to regulate the learning process by 1) minimizing expression classification error, 2) minimizing intra-class variation within the same expression class, 3) minimizing expression-state classification error, 4) minimizing expression-state variation, and 5) preserving the expression-state continuity, which is necessary for consecutive temporal-feature representation learning in the second part.

In order to capture the facial expression dynamics, temporal feature representation of the facial expression is learned via the LSTM of the second part. The CNN model learned in the first part is utilized to extract spatial feature representation of each frame for training the LSTM. The expression class separability of the learned spatial feature representation is improved due to the expression-state constraints. On top of the spatial feature representation, the LSTM can consecutively learn the temporal dynamics of the facial expression. As a result, the proposed method can generate discriminative spatio-temporal feature representations that improve the recognition of facial expression at different expression intensities. The details of each step of the proposed method are described in the following subsections.

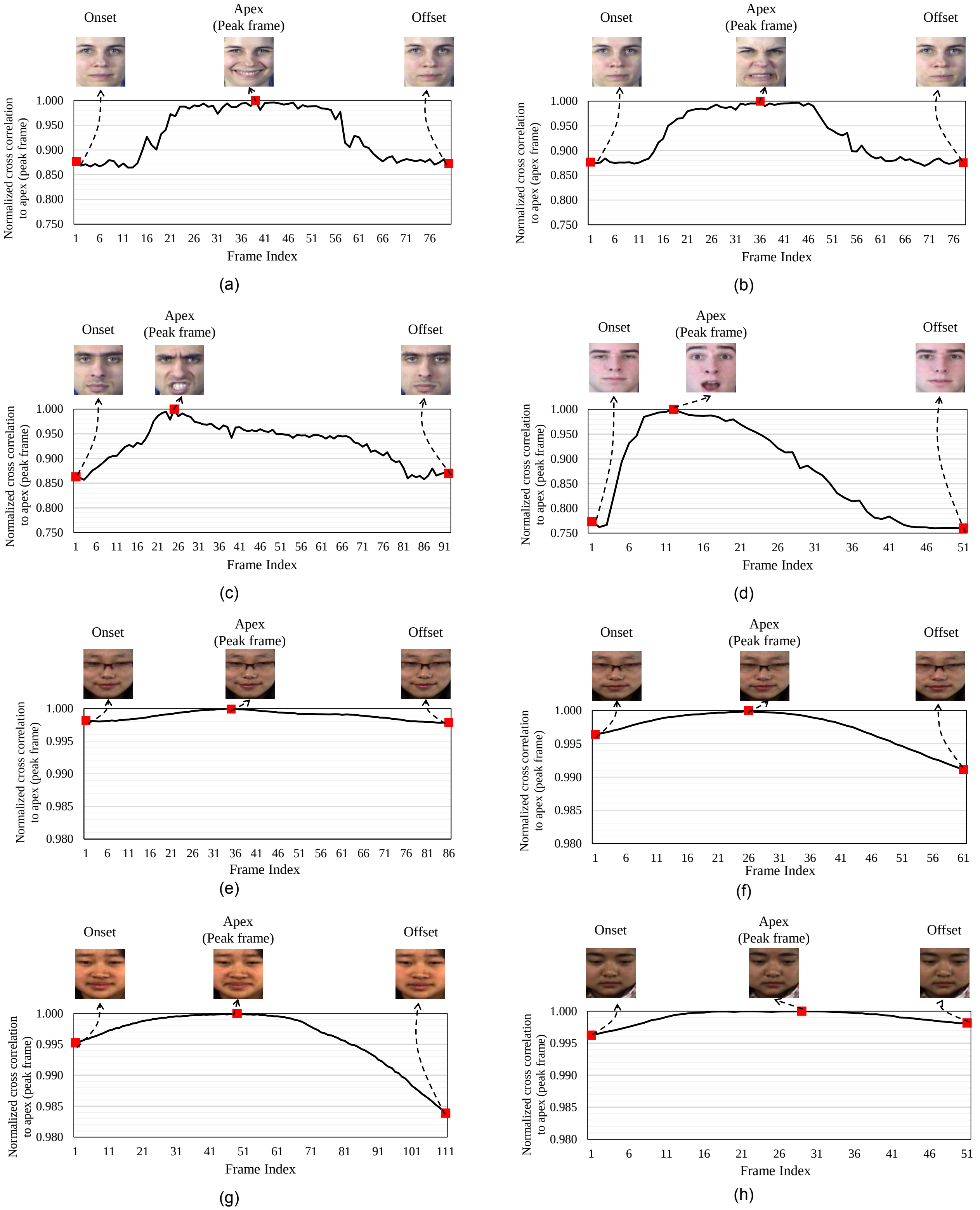


Fig. 2. Examples of the change of expression state, which is represented as normalized cross correlations between apex (peak frame) and onset, apex and offset frames. (a) Happy, (b) disgust, (c) angry and (d) surprise expressions from MMI dataset. (e) Happy, (f) disgust, (g) repression and (h) surprise expressions from CASME II dataset.

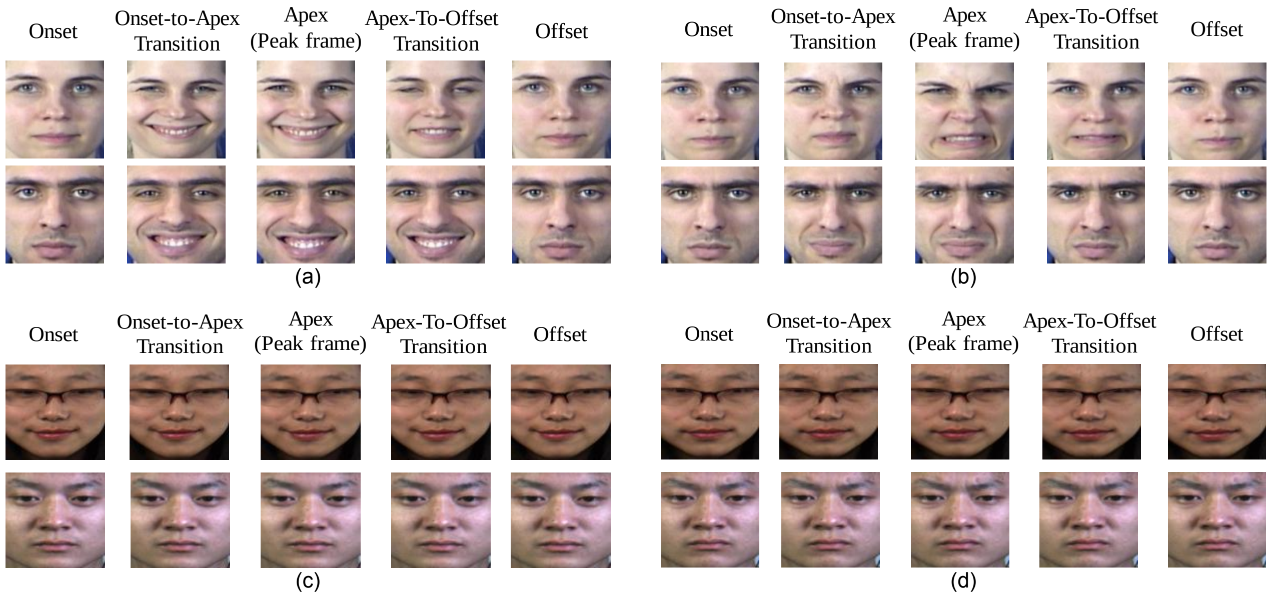


Fig. 3. Examples of the extracted representative expression-state frames. (a) Happy and (b) disgust expressions from MMI dataset. (c) Happy and (d) disgust expressions from CASME II dataset.

TABLE 1
THE ARCHITECTURE OF THE CNN FOR SPATIAL FEATURE REPRESENTATION LEARNING WITH EXPRESSION-STATE CONSTRAINTS

Layer*	Input shape	Kernel	Stride	Pad	Output shape
Input	64×64×3	-	-	-	64×64×3
C1	64×64×3	3×3	1	0	62×62×32
S1	62×62×32	3×3	2	0	31×31×32
C2	31×31×32	3×3	1	0	29×29×64
S2	29×29×64	3×3	2	0	14×14×64
C3	14×14×64	3×3	1	0	12×12×64
S3	12×12×64	3×3	2	0	6×6×64
F4	6×6×64	-	-	-	512×1
F5 **	512×1	-	-	-	512×1
F_c ***	512×1 (output of F5)	-	-	-	N_c×1
F_p ***	512×1 (output of F5)	-	-	-	N_p×1

* *C, S and F are abbreviations for convolutional layer, max pooling layer and fully connected layer, respectively.*

** *The spatial feature representation is extracted from the layer* F5.

*** *The two layers* F_c *and* F_p *are only used for training the CNN. And* N_c *is the number of expression classes and* N_p *is the number of expression-states.*

2.2 Spatial Feature Representation Learning with Expression-State Constraints

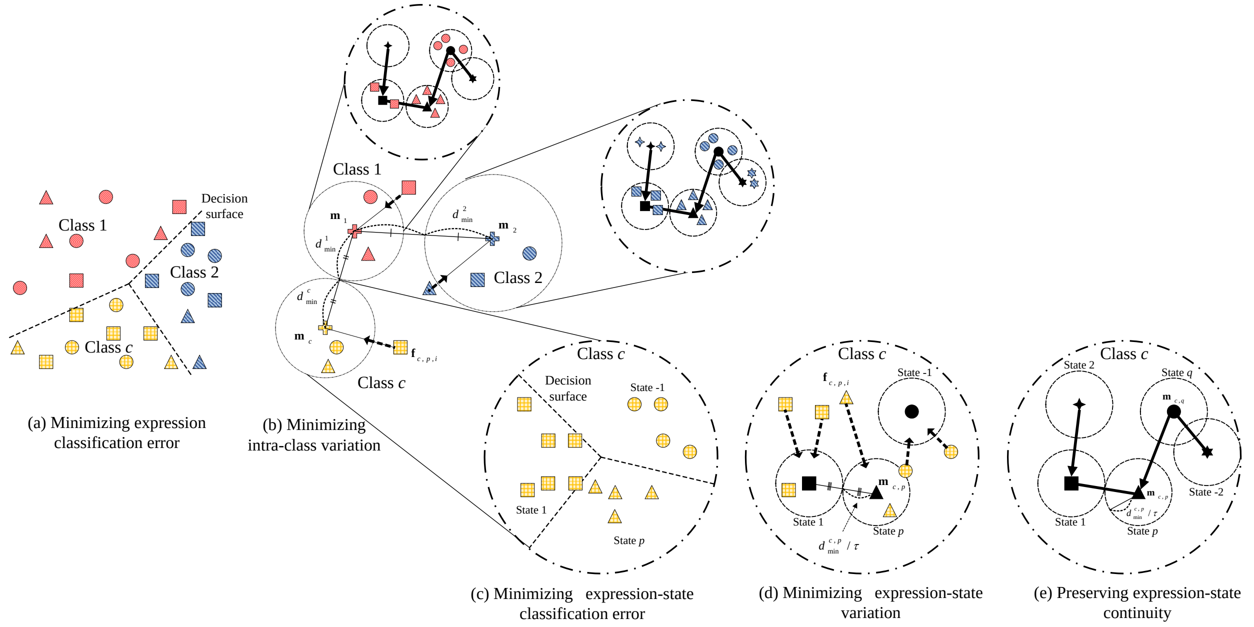
2.2.1 Representative Expression-State Frame Extraction

In practice, facial expressions appear in a wide range of expression intensities. The expression intensities can be measured by determining deformation of facial parts, which result in difference of facial images [38]. The facial expression intensity can be measured by calculating correlation between features of the reference frame and other frames [39]. In this paper, normalized cross correlation [40] of each frame with respect to the apex frame is used to measure the changes between facial frames, which is a widely used image similarity metric due to its robustness to linear changes in the amplitude of illumination [41]. Fig. 2 shows examples of the change of expression state, which is represented by the normalized cross correlations. As shown in the figure, the expression intensity varies in a wide range across subjects and expression classes. Such expression intensity variations make the practical FER a challenging problem [14]. To overcome this challenge, representative expression-states are utilized, which are specified in facial sequences regardless of the expression intensity variations and different expression duration [35], [36]. Representative expression-state frames are selected from an expression sequence as

$$\mathbf{I}' = \left\{ \mathbf{I}_t \,\middle|\, t = \left[t_{on}, \frac{t_{on}+t_{apex}}{2}, t_{apex}, \frac{t_{apex}+t_{off}}{2}, t_{off} \right] \right\}, \tag{1}$$

where $\mathbf{I}'$ is a set of representative expression-state frames, $\mathbf{I}_t$ is the t-th frame. t_{on}, t_{apex} and t_{off} are indexes of onset, apex and offset frames provided by the ground truth, respectively. As a result, one expression sequence is represented with five representative expression-state frames, namely, onset, onset-to-apex transition, apex, apex-to-offset transition and offset. The selected representative expression-state frames are used to train the CNN model of the proposed framework. Examples of the extracted representative expression-state frames are shown in Fig. 3.

By utilizing the expression-state as a learning constraint of the CNN, the proposed method improves expression class separability of the learned spatial feature representation. In addition, the use of the representative expression-states enables the spatial feature representation learning to handle a variety of expression transition duration.



(a) Minimizing expression classification error

(b) Minimizing intra-class variation

(c) Minimizing expression-state classification error

(d) Minimizing expression-state variation

(e) Preserving expression-state continuity

Fig. 4. Illustration of the objective terms of the proposed objective function for the proposed spatial feature representation learning with expression-state constraints. Color represents the expression class of samples and shape represents expression-state of samples.

2.2.2 Learning Spatial Feature Representation with the Proposed Objective Function Constraints to Expression-States

Table 1 shows the architecture used for the spatial feature representation learning with expression-state constraints. The architecture is based on LeNet architecture [42]. The parameters on the architecture in Table 1 were experimentally chosen for computational efficiency and classification accuracy. Note that in the network, F$_c$ layer is used for expression classification and F$_p$ layer is used for expression state classification. After training the network, the proposed spatial feature representation is extracted at the layer F5. To improve the expression class separability of the spatial feature representation during the training of the CNN, five objective terms illustrated in Fig. 4 are devised. In addition, gradients of the objective terms used for back-propagation algorithm [43], [44] are accordingly derived. Note that Table B in Appendix B summarizes variables used in the proposed objective terms.

The first objective term (E_1) of the objective function (Fig. 4 (a)) is devised for minimizing the expression classification error at the layer F$_c$. To that end, a cross entropy error [45] can be readily utilized to enforce the learning of discriminative feature representation, which is defined as

$$E_1 = -\sum_i \sum_k y^{(F_c)}_{i,k} \log \hat{y}^{(F_c)}_{i,k}, \qquad (2)$$

where $y^{(F_c)}_{i,k}$ is the expression ground truth of the i-th sample (1 if k is the correct class and 0 otherwise), and $\hat{y}^{(F_c)}_{i,k}$ is the predicted probability that the sample belongs to the expression class k calculated at the layer F$_c$ with Softmax function. The objective term E_1 makes samples with different expression classes be separable in the feature space as shown in Fig. 4 (a).

To perform back-propagation, the gradients of the objective term E_1 for the layer F$_c$ are written as follows (please see Appendix A for the detailed derivation):

$$\frac{\partial E_1}{\partial \mathbf{b}^{(F_c)}} = \sum_i \left(\hat{\mathbf{y}}^{(F_c)}_i - \mathbf{y}^{(F_c)}_i\right) \text{ and } \frac{\partial E_1}{\partial \mathbf{W}^{(F_c)}} = \mathbf{h}^{(F5)}\left(\frac{\partial E_1}{\partial \mathbf{b}^{(F_c)}}\right)^T, \qquad (3)$$

where $\mathbf{y}^{(F_c)}_i = [y^{(F_c)}_{i,1}, y^{(F_c)}_{i,2}, ..., y^{(F_c)}_{i,k}, ..., y^{(F_c)}_{i,N_c}]^T$ is the expression ground truth vector of the i-th training sample, $y^{(F_c)}_{i,k} = 1$ if and only if the true class of the sample is k-th class, and $\hat{\mathbf{y}}^{(F_c)}_i$ denotes the predicted probability calculated at the layer F$_c$ with Softmax function. $\mathbf{b}^{(F_c)}$ and $\mathbf{W}^{(F_c)}$ denote the biases and weights of the layer F$_c$, respectively. $\mathbf{h}^{(F5)}$ denotes the output of the last hidden layer (F5 in Table 1). As shown in (3), the gradient with respect to the weight is dependent on the gradient with respect to the bias. Therefore, the gradients with respect to the bias are sufficient to be described in the objective term. The gradients of the other layers can also be computed by the same back-propagation algorithm [43], [44].

In addition, the intra-class variations in the same expression class could occur due to factors such as the subject appearance. To mitigate those intra-class variations of spatial feature representations at the last hidden layer (F5 in Table 1), the term E_2 (Fig. 4 (b)) is devised as

$$E_2 = \frac{1}{2}\sum_{c,p,i} g\left(\left\|\mathbf{f}_{c,p,i} - \mathbf{m}_c\right\|_2^2 - (d^c_{\min})^2\right), \qquad (4)$$

where $\mathbf{f}_{c,p,i}$ is the spatial feature representation vector of the i-th training sample of class c and the p-th expression-state extracted at the last hidden layer (F5 in Table 1), $\mathbf{m}_c$ is the mean feature vector of the training samples in class c and $d^c_{\min}$ is half of the minimum distance between $\mathbf{m}_c$ and $\mathbf{m}_j$ for $j \neq c$. The function $g(\omega) = \log(1 + \exp(\beta\omega))/\beta$ is a smoothed approximation of $[\omega]_+ = \max(0, \omega)$ [46] where β is a sharpness parameter. As shown in Fig. 4 (b), the minimization of (4) makes $\mathbf{f}_{c,p,i}$ be in a multi-dimensional sphere which radius is $d^c_{\min}$ and the center point is $\mathbf{m}_c$, thus resulting in the minimization of intra-class variation.

For simplicity when obtaining the gradient of the second objective term (E_2) in (4), the mean feature vector $\mathbf{m}_c$ is assumed to be a constant vector. Thus the gradient of E_2

with respect to the bias of the last hidden layer (F5 in Table 1) can be written as

$$\frac{\partial E_2}{\partial \mathbf{b}^{(F5)}} = \sum_{c,p,i} g'(e_2)(\mathbf{f}_{c,p,i} - \mathbf{m}_c) \circ \sigma'(\mathbf{h}^{(F5)}),$$
$$e_2 = \left\|\mathbf{m}_c - \mathbf{f}_{c,p,i}\right\|_2^2 - (d_{\min}^{c})^2, \tag{5}$$

where '$\circ$' denotes the Hadamard product, $\mathbf{b}^{(F5)}$ denotes the bias of the last hidden layer (F5 in Table 1), $\sigma(\cdot)$ is the activation function, and $\mathbf{h}^{(F5)} = \mathbf{W}^{(F5)}\mathbf{h}^{(F4)} + \mathbf{b}^{(F5)}$.

As can be observed from Fig. 3, expression-state frames could appear similar to other expression-state frames, due to small facial motion changes. Such similarity between the frames could result in similar spatial feature representation along the sequence of frames. As a result, temporal changes become harder to model, especially in micro-expression sequences. To improve the distinction between the learned spatial feature representations of different expression-state frames, two objective terms (E_3 and E_4) are devised. E_3 (Fig. 4 (c)) is added to distinguish the expression-state by minimizing the expression-state classification error at the layer F$_p$ as

$$E_3 = -\sum_i \sum_j y_{i,j}^{(F_p)} \log\, \hat{y}_{i,j}^{(F_p)}, \tag{6}$$

where j is the expression-state index, $y_{i,j}^{(F_p)}$ is the expression-state ground truth of the i-th sample (1 if j is the correct expression-state and 0 otherwise), and $\hat{y}_{i,j}^{(F_p)}$ is the predicted expression-state of that sample calculated at the layer F$_p$ with Softmax function which generates predicted probability of each expression-state. The term E_3 makes samples with different expression-states in an expression class be separable in the feature space as shown in Fig. 4 (c).

Similar to (3), the gradients of the third objective term (E_3) are given by

$$\frac{\partial E_3}{\partial \mathbf{b}^{(F_p)}} = \sum_i \left(\hat{\mathbf{y}}_i^{(F_p)} - \mathbf{y}_i^{(F_p)}\right) \text{ and } \frac{\partial E_3}{\partial \mathbf{W}^{(F_p)}} = \mathbf{h}^{(F5)}\left(\frac{\partial E_3}{\partial \mathbf{b}^{(F_p)}}\right)^T, \tag{7}$$

where $\mathbf{y}_i^{(F_p)} = [y_{i,1}^{(F_p)}, y_{i,2}^{(F_p)}, \ldots, y_{i,j}^{(F_p)}, \ldots, y_{i,N_p}^{(F_p)}]^T$ is the expression-state ground truth vector of the i-th training sample. $y_{i,j}^{(F_p)} = 1$ if and only if the true expression-state of the sample is j-th expression-state. $\hat{\mathbf{y}}_{i,j}^{(F_p)}$ denotes the predicted probability calculated at the layer F$_p$. $\mathbf{b}^{(F_p)}$ and $\mathbf{W}^{(F_p)}$ denote the biases and weights of the layer F$_p$, respectively.

As shown in Fig. 3, even same expression-state of the expression class could contain appearance variations (e.g., subject appearance variations, expression intensity variations, etc.). To minimize the variations of expression-state in spatial feature representations at the last hidden layer (F5 in Table 1), the term E_4 (Fig. 4 (d)) is devised as

$$E_4 = \frac{1}{2}\sum_{c,p,i} g\left(\left\|\mathbf{f}_{c,p,i} - \mathbf{m}_{c,p}\right\|_2^2 - \left(\frac{d_{\min}^{c,p}}{\tau}\right)^2\right), \tag{8}$$

where $\mathbf{m}_{c,p}$ is the mean feature vector of samples in the c-th class and the p-th expression-state and $d_{\min}^{c,p}$ is half of the minimum distance between $\mathbf{m}_{c,p}$ and $\mathbf{m}_{c,j}$ for $j \neq p$. τ is a parameter for determining the range of the expression-state distribution. As shown in Fig. 4 (d), the minimization of (8) makes $\mathbf{f}_{c,p,i}$ be in a multi-dimensional sphere which radius is $d_{\min}^{c,p}$ and the center point is $\mathbf{m}_{c,p}$. As a result, it helps cluster the spatial feature representation of the same expression-state, which minimizes the effect of the expression-state variations on the FER task.

By considering the feature means $\mathbf{m}_{c,p}$ as constant vectors, the gradient of E_4 with respect to bias can be written as

$$\frac{\partial E_4}{\partial \mathbf{b}^{(F_p)}} = \sum_{c,p,i} g'(e_4)(\mathbf{f}_{c,p,i} - \mathbf{m}_{c,p}) \circ \sigma'(\mathbf{h}^{(F5)}),$$
$$e_4 = \left\|\mathbf{m}_{c,p} - \mathbf{f}_{c,p,i}\right\|_2^2 - \left(\frac{d_{\min}^{c,p}}{\tau}\right)^2. \tag{9}$$

The terms E_3 and E_4 are devised to improve the distinction between expression-states. However, spatial feature representations from adjacent frames in an expression sequence are not guaranteed to be continuous to each other in the feature space. The continuity of spatial feature representations in adjacent frames is useful to expression sequence classification in the following temporal feature representation learning (described in section 2.3). Thus, the last term, E_5 (Fig. 4 (e)), is devised to regulate the expression-state continuity in the feature space. As a result, spatial feature representations of frames between adjacent expression-states are enforced to reside in between the spatial feature representations in the feature space. The term E_5 is defined as

$$E_5 = \frac{1}{2}\sum_c \sum_{p,p \neq p_{apex}} g\left(\left\|\mathbf{m}_{c,q} - \mathbf{m}_{c,p}\right\|_2^2 - \left(\frac{2d_{\min}^{c,p}}{\tau}\right)^2\right),$$
$$q = \begin{cases} p+1, & \text{if } p < p_{apex} \\ p-1, & \text{if } p > p_{apex} \end{cases}, \tag{10}$$

where p_{apex} is the index of apex expression-state, q represents the index of the adjacent expression-state in the learning stage. As shown in Fig. 4 (e), the term E_5 is devised to feature means of two adjacent expression-states $\mathbf{m}_{c,p}$ and $\mathbf{m}_{c,q}$ become continuous to each other in the feature space.

Similar to (9), by considering the feature means $\mathbf{m}_{c,p}$ and $\mathbf{m}_{c,q}$ as constant vectors, the gradients of E_5 with respect to bias of the last hidden layer (F5 in Table 1) can be written as

$$\frac{\partial E_5}{\partial \mathbf{b}^{(L)}} = \sum_{c,p,p \neq 0} \frac{g'(e_5)}{N_{c,p}}(\mathbf{m}_{c,p} - \mathbf{m}_{c,q}) \circ \sigma'(\mathbf{h}^{(F5)}),$$
$$e_5 = \left\|\mathbf{m}_{c,q} - \mathbf{m}_{c,p}\right\|_2^2 - \left(\frac{2d_{\min}^{c,p}}{\tau}\right)^2, \tag{11}$$

where $N_{c,p}$ is the number of samples from all subjects with expression-states for the training set.

To validate the improvement of the expression class separability according to the proposed objective terms, Fisher's discriminant ratio [47] is employed as a class separability score. The Fisher's discriminant ratio (J) measures the ratio of the between-class scatter (S_b) and the within-class scatter (S_w) as follows

$$J = \frac{trace\ (S_b)}{trace\ (S_w)}, \tag{12}$$

$$S_b = \sum_{i=1}^{N_c} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \quad S_w = \sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\mathbf{f}_{ij} - \mathbf{m}_i)(\mathbf{f}_{ij} - \mathbf{m}_i)^T,$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2695999, IEEE Transactions on Affective Computing

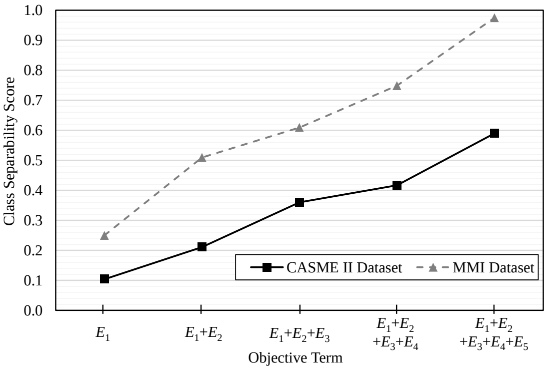


Fig. 5. Comparison of class separability score of the spatial feature representation learned with different objective terms in MMI dataset and CASME II dataset. Note that higher value means more separable classes.

TABLE 2

THE ARCHITECTURE OF THE NETWORK FOR LEARNING THE SPATIO-TEMPORAL FEATURE REPRESENTATION OF THE EXPRESSION DYNAMICS

Type	Layer*	Input shape	Output shape
Input	Input	$N_f \times (64{\times}64{\times}3)$	$N_f \times (64{\times}64{\times}3)$
Spatial feature representation	C1~F5 in Table 1	$N_f \times (64{\times}64{\times}3)$	$N_f \times (512{\times}1)$
Temporal feature representation	L6	$N_f \times (512{\times}1)$	$N_f \times (512{\times}1)$
	L7	$N_f \times (512{\times}1)$	$N_f \times (512{\times}1)$
	L8	$N_f \times (512{\times}1)$	$N_f \times (512{\times}1)$
	F9	$N_f \times (512{\times}1)$	$N_c \times 1$

* *L is an abbreviation of LSTM layer.*

N_f is the number of frames and N_c is the number of expression classes.

where N_c is the number of classes, N_i is the number of samples in the i-th expression class, $\mathbf{m}_i$ is the mean feature vector of samples in the i-th class, $\mathbf{m}$ is the mean feature vector of all samples and $\mathbf{f}_{ij}$ is the spatial feature representation of the j-th sample in the i-th class. The spatial feature representation was extracted from the last hidden layer of the trained CNN (F5 in Table 1). Higher Fisher's discriminant ratio is achieved with larger between-class scatter and smaller within-class scatter which results in better expression class separability.

Fig. 5 shows the class separability scores with different objective terms in the objective function. As can be seen in the figure, by employing E_2, which focuses on reducing the intra-class variations of each expression, the class separability score in feature space is improved. Significant improvement is achieved by including terms related to the expression-state (i.e., E_3, E_4 and E_5) in the objective function.

2.3 Temporal Feature Representation Learning with LSTM

In this section, temporal feature representation learning with LSTM is described. LSTM is employed to model facial dynamics of the variable length facial expression sequences. Table 2 summarizes the overall architecture of the proposed network. Motivated by the recent deep learning approaches [48], [49], deep LSTM network is constructed by stacking multiple LSTM layers on top of each other. This stacked LSTM network can combine multiple representations with flexible use of face sequences [48]. In the network, the spatial feature representations of all the frames $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{N_f}]$ extracted from the learned CNN model are fed to the LSTM. As a result, the last LSTM layer (L8) produces the proposed spatio-temporal feature representation. Then, the classification result is drawn at F9 with a Softmax classifier.

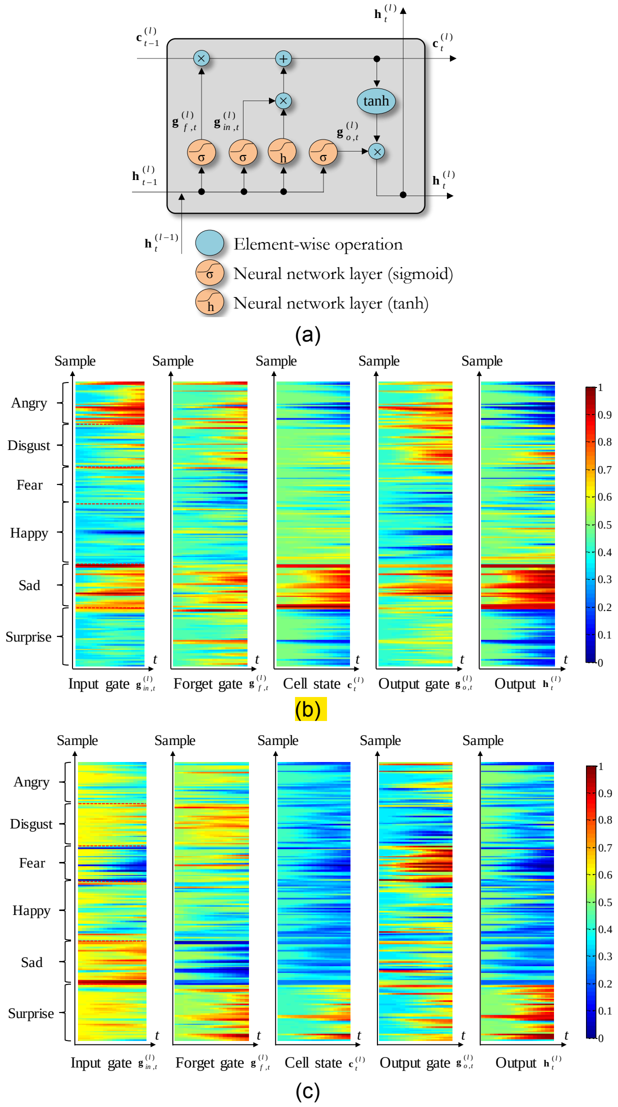


Fig. 6. (a) Illustration of the LSTM (defined in (13)) used for temporal feature representation learning. (b) and (c) Corresponding visualization of gates, cell state and output of last LSTM layer (i.e., l=L8). The horizontal axis t corresponds to the number of processed frames. The vertical axis is the index of the video sequence samples in the MMI dataset. The number of frames in x-axis and the number of samples in y-axis of the figure are 20 and 197, respectively. Here we visualize the gates and cell state of two LSTM dimensions that successfully learn patterns of (b) sad expression and (c) surprise expression in temporal domain, respectively. As shown in (a), input gate ($\mathbf{g}_{in,t}^{(l)}$) with large values stores output $\mathbf{h}_t^{(l-1)}$ from the lower LSTM layer ((l-1)-th layer) to the cell state ($\mathbf{c}_t^{(l)}$). Forget gate ($\mathbf{g}_{f,t}^{(l)}$) with large values retrains the information in the cell state ($\mathbf{c}_t^{(l)}$). Output gate ($\mathbf{g}_{o,t}^{(l)}$) with large values generates a latent feature representation from the cell state ($\mathbf{c}_t^{(l)}$) that includes expression class related information.

Fig. 6 (a) illustrates the structure of the LSTM defined in (13). Fig. 6 (b) and (c) visualize the evolution of the states of three gates, a memory cell and an output (i.e., latent feature representation) for two LSTM dimensions that are activated on sad expression (Fig. 6 (b)) or a surprise expression (Fig. 6 (c)) at the last LSTM layer L8 in the MMI dataset. As shown in the figures, the input gate determines which information should be added to the memory cell. The forget gate decides which information stored in the memory cell is important and should be retained (i.e., larger values are activated at the forget gate to retain information in its memory cell). As the network processes more frames, the memory cell states gradually absorb the useful information related to expression classes. And the output gate makes a latent feature representation of the output data related to expression classes. Note that the LSTM layer illustrated in Fig. 6 (a) operates as follows [37]

$$\begin{aligned}
\mathbf{g}_{in,t}^{(l)} &= \mathrm{sigm}(\ \mathbf{W}_{in}^{(l)}[\mathbf{h}_{t-1}^{(l)},\mathbf{h}_{t}^{(l-1)}] + \mathbf{b}_{in}^{(l)}),\\
\mathbf{g}_{f,t}^{(l)} &= \mathrm{sigm}(\ \mathbf{W}_{f}^{(l)}[\mathbf{h}_{t-1}^{(l)},\mathbf{h}_{t}^{(l-1)}] + \mathbf{b}_{f}^{(l)}),\\
\mathbf{g}_{o,t}^{(l)} &= \mathrm{sigm}(\ \mathbf{W}_{o}^{(l)}[\mathbf{h}_{t-1}^{(l)},\mathbf{h}_{t}^{(l-1)}] + \mathbf{b}_{o}^{(l)}),\\
\mathbf{c}_{t}^{(l)} &= \mathbf{c}_{t-1}^{(l)} \circ \mathbf{g}_{f,t}^{(l)} + \tanh(\ \mathbf{W}_{c}^{(l)}[\mathbf{h}_{t-1}^{(l)},\mathbf{h}_{t}^{(l-1)}] + \mathbf{b}_{c}^{(l)}) \circ \mathbf{g}_{in,t}^{(l)},\\
\mathbf{h}_{t}^{(l)} &= \tanh(\ \mathbf{c}_{t}^{(l)}) \circ \mathbf{g}_{o,t}^{(l)},
\end{aligned} \tag{13}$$

where $\mathbf{W}_{*}^{(l)}$ and $\mathbf{b}_{*}^{(l)}$ denotes the weights and biases of the l-th LSTM layer. $\mathbf{g}_{in,t}^{(l)}$ is an input gate, $\mathbf{g}_{f,t}^{(l)}$ is a forget gate, $\mathbf{g}_{o,t}^{(l)}$ is an output gate, $\mathbf{c}_{t}^{(l)}$ is the memory cell, and $\mathbf{h}_{t}^{(l)}$ is the output of the l-th LSTM layer with a given t-th input.

3 Experiments

3.1 Experimental Setup

To verify the effectiveness of the proposed FER with learned spatio-temporal feature representation, experiments have been conducted on the MMI dataset [36] for deliberate expressions, and the CASME II dataset [35] for micro-expressions. In the experiments, the face region in each frame was automatically cropped and aligned based on the eye landmarks [50] which were automatically detected [51]. The construction of the utilized MMI and CASME II datasets was performed as follows

1. MMI dataset: A total of 205 deliberate expression sequences with frontal faces were collected from 30 subjects. The expression sequences were recorded at a temporal resolution of 24 fps. Each expression sequence of the dataset was labeled with one of the six basic expression classes (i.e., angry, disgust, fear, happy, sad, and surprise). The expression sequences were collected such that, the first frame in the sequence was the onset frame and last frame was the offset frame. The indexes of the apex frames were located manually [14].
2. CASME II dataset: A total of 246 spontaneous micro-expression sequences were collected from 26 subjects. The expression sequences were recorded at a temporal resolution of 200 fps. The micro-expression sequences in the dataset were labeled as one of the five expression classes (i.e., happiness, disgust, repression, surprise, and others). The onset, apex and offset frame indexes were also provided for each micro-expression sequence. In the experiments, micro-expression sequences from onset-to-offset were used [52].

All the experiments in this paper were conducted in a subject independent manner, such that the subjects in the training set were excluded from the test set. Particularly, the experiments were conducted with a leave-one-subject-out (LOSO) cross validation [3], [19], [24]. To avoid overfitting due to the limited number of samples in the utilized datasets (MMI and CASME II), expression sequences were augmented during the training [53]. For each training expression sequence, 150 augmented sequences were obtained by: 1) horizontal flipping of the sequence frames, 2) rotating the frames between the angles [-10°, 10°] with an increment of 5°, 3) translating the frames along [(0, 0), (-2, -2), (-2, +2), (+2, -2), (+2, +2)] pixels in x and y axis, and 4) scaling the frames with scaling factors of 0.9, 1.0 and 1.1. As a result, about 29k and 35k training sequences were generated for each cross validation fold on average in MMI and CASME II datasets, respectively. For training the CNN, 5 representative expression-state frames were selected from each augmented training sequence as described in section 2.2.1, resulting in 148k and 177k training frames for each cross validation fold on average in MMI and CASME II datasets, respectively. After that, the learned CNN model was utilized to extract the spatial feature representation of all frames of training sequences for training the LSTM. The learning of the CNN (shown in Table 1) with the proposed objective terms was implemented by ConvNet [54]. The temporal feature representation learning with LSTM (Table 2) was implemented by Keras [55]. For the activation function, rectified linear unit (ReLU) [56] was used. In this paper, the initial learning rate was set to 0.08 for the MMI dataset and 0.01 for the CASME II dataset. The training epochs were set to 40 and 50 for the CNN and LSTM, respectively. The training of the CNN and LSTM for the given number of epoch took approximately 150 minutes and 100 minutes on a NVIDIA Titan X GPU, respectively.

3.2 Effectiveness of the Proposed Spatio-Temporal Feature Representation

We demonstrated the effectiveness of the proposed spatio-temporal feature representation for the FER task under a subject-independent recognition scenario. For the MMI dataset, the comparison was conducted with state-of-the-art methods including hand-crafted features and deep learning. As hand-crafted feature based methods, sparse representation classifier (SRC) [3], [24] with static (frame-based) features (i.e., LBP [8], LPQ [10], Gabor wavelet [9], and collaborative expression representation (CER) [3] extracted from the apex frames) and spatio-temporal features (i.e., LBP-TOP [12] and LPQ-TOP [11, 13]) were used. In addition, deep learning based methods with spatial feature representation (i.e., AURF [30], AUDN [31], and deep CNN [26]) and spatio-temporal feature representation (i.e., 3D CNN-DAP [19] and DTAGN [32]) were used in the comparison.

TABLE 3
RECOGNITION PERFORMANCE WITH MMI DATASET

(a)

Type	Method	Cross validation	Input	Recognition rate (%)
Hand-crafted feature based methods	LBP + SRC [8]	LOSO	Apex frame	59.18
	LPQ + SRC [10]	LOSO	Apex frame	62.72
	Gabor + SRC [9]	LOSO	Apex frame	61.89
	LBP-TOP [12]+SRC	LOSO	Sequence	61.19
	LPQ-TOP [11], [13] + SRC	LOSO	Sequence	64.11
	CER [3]	LOSO	Apex frame	70.12
Deep learning based methods	3D CNN-DAP [19]	LOSO	Sequence	63.40
	DTAGN [32]	10-fold	Sequence	70.24
	AURF [30]	10-fold	Apex frame	69.88
	AUDN [31]	10-fold	Apex frame	75.85
	Deep CNN [26]	5-fold	Static frame	77.90
	Proposed spatial feature representation	LOSO	Apex frame	69.94
	Proposed spatio-temporal feature representation	LOSO	Sequence	**78.61**

(b)

Predicted Actual	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	**77.78**	14.81	0.00	3.70	3.70	0.00
Disgust	4.17	**83.33**	4.17	4.17	4.17	0.00
Fear	0.00	0.00	**65.22**	13.04	8.70	13.04
Happy	0.00	3.03	0.00	**93.94**	0.00	3.03
Sad	10.71	0.00	10.71	3.57	**71.43**	3.57
Surprise	5.26	10.53	2.63	2.63	2.63	**76.32**

(a) Comparisons with state-of-the-art FER methods. (b) Confusion matrix of the proposed method.

TABLE 4
RECOGNITION PERFORMANCE WITH CASME II DATASET

(a)

Method	Recognition rates (%)
LBP-TOP [12]	44.12
LBP-TOP with adaptive motion magnification [52]	51.91
LBP-MOP [15]	45.75
Riesz wavelet [57]	46.15
Proposed spatial feature representation	58.54
Proposed spatio-temporal feature representation	**60.98**

(b)

Predicted Actual	Happy	Disgust	Surprise	Repression	Others
Happy	**37.50**	12.50	3.13	0.00	46.88
Disgust	1.59	**50.79**	3.17	0.00	44.44
Surprise	0.00	8.00	**60.00**	0.00	32.00
Repression	14.81	7.41	3.70	**22.22**	51.85
Others	0.00	12.12	0.00	2.02	**85.86**

(a) Comparisons with state-of-the-art FER methods. (b) Confusion matrix of the proposed method.

The comparative recognition rates and the confusion matrix of the proposed method are shown in Table 3. As shown in the table, the proposed method outperformed existing state-of-the-art FER methods. Specifically, the proposed spatio-temporal feature representation showed better recognition rates compared to the deep learning based methods with spatial feature representation (AURF [30], AUDN [31], deep CNN [26]) and the proposed spatial feature representation. This can be attributed to the efficient encoding of the expression dynamics in the proposed spatio-temporal feature representation learning. Moreover, by utilizing the expression-state in the proposed objective function, the proposed method achieved a superior performance to methods that encoded the spatio-temporal dynamics with a 3D CNN (3D CNN-DAP [19] and DTAGN [32]).

In the spontaneous micro-expression experiments, the FER performance of the proposed method was compared with four spatio-temporal hand-crafted feature based methods devised for the micro-expression recognition; 1) Support vector machine (SVM) classifier with LBP-TOP [12]; 2) SVM with LBP-TOP extracted from motion magnified sequences [52]; 3) SVM with LBP-Three Mean Orthogonal Planes (LBP-MOP) [15]; and 4) SVM with Monogenic Riesz wavelet representation [57].

The comparative recognition rates and the confusion matrix of the proposed method on the CASMI II dataset are shown in Table 4. As shown in the table, the proposed spatio-temporal feature representation outperformed existing state-of-the-art micro-expression recognition methods and the proposed spatial feature representation. The results indicate that the proposed method is also effective for recognizing micro-expressions.

3.3 Inter-Dataset Evaluation of the Proposed Spatio-temporal Feature Representation

To investigate the generalization performance of the proposed method, an inter-dataset evaluation was conducted. In deep learning, it is known that lower layers normally extract general low-level features such as edges from the input images while higher layers extract high-level features which are more specific to different tasks [58]. Thus, it can be expected that the network that learned subtle facial dynamic patterns on micro-expression could operate well on recognizing deliberate expression that includes apparent facial dynamics. For this experiment, the network was trained on the micro-expression dataset (CASME II) and tested on the deliberate expression dataset (MMI). In training the network, due to the different expression classes in both datasets, fine tuning was performed in order to represent all the expressions in the test dataset (MMI). To that end, weights of the first two convolutional layers (C1 and C2) were initialized with pre-trained weights trained from CASME II dataset. Remaining higher layers were initialized by the Xavier's initialization [59]. During the training of the network, weights of the first two convolutional layers (C1 and C2) were fixed, while higher layers (from C3 to F9) were trained. In this experiment, leave-one-subject-out cross validation scheme was also used, as described in section 3.1.

TABLE 5

RECOGNITION RATES OF THE PROPOSED METHOD UNDER INTRA AND INTER-DATASET EVALUATION

Evaluation method (Training set / test set)	Recognition rates (%)
Intra-dataset (MMI / MMI)	78.61
Inter-dataset (CASME II / MMI)	72.83

Table 5 shows the recognition rates of the proposed method under intra and inter-dataset evaluation. As shown in the table, the proposed method was able to maintain a feasible recognition rate under inter-dataset evaluation. Furthermore, the proposed method under inter-dataset evaluation showed comparable performance to state-of-the-art methods under intra-dataset evaluation shown in Table 4. This result indicates that the proposed method could learn versatile spatio-temporal feature representation of micro-expression and deliberate expression.

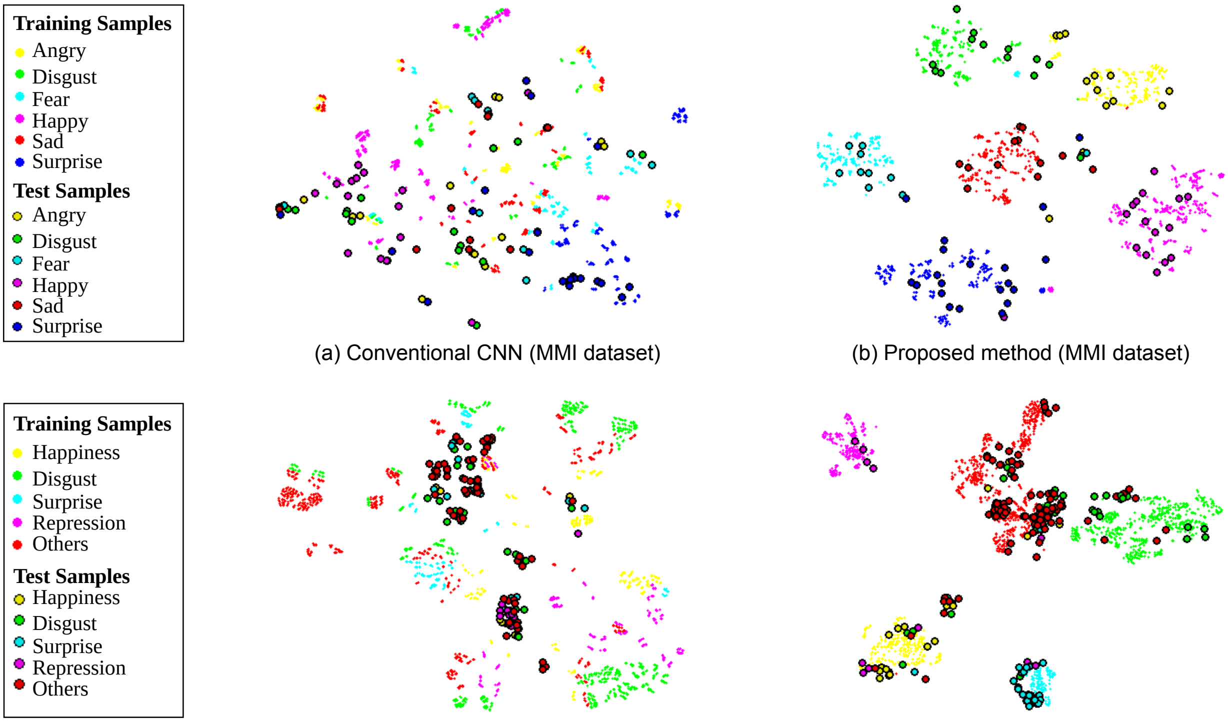
3.4 Effectiveness of the Proposed Spatial Feature Representation Learning with Expression-State Constraints

In this section, the effects of the proposed objective terms were evaluated in terms of the expression recognition rate in the CNN (the part 1 in Fig. 1). For the evaluation, the proposed objective terms were incrementally added to the objective function for improving the class separability of the feature space in a coarse to fine approach. First, expression class related objective terms (i.e., E_1 and E_2) were used, because the primary goal is the expression classification. Then, expression-state related objective terms (i.e., E_3 and E_4) were followed. If expression-state related objective terms are used first, the training may suffer from local minima, because expression-state class could be fine-grained class of expression class. In addition, minimizing intra-class variation related terms (i.e., E_2 and E_4) should be followed by minimizing classification error related terms (i.e., E_1 and E_3), because minimizing intra-class variation does not guarantee minimization of classification error. Lastly, preserving the expression-state continuity term (i.e., E_5) should be followed by expression-state related terms (i.e., E_3 and E_4). For applying the expression-state continuity term, expression-states need to be differentiable in feature space. In case of the spatial feature representation which is a static (frame-based) feature extracted from a frame, apex frame has the most information for the expression classification. Therefore, the performance was measured using apex frames of each sequence. The recognition rates were calculated with an output of layer Fc which predicts the probability of each expression class. The t-test was employed to measure the statistical significance of the recognition rate improvement of each objective term. For the t-test, the leave-one-subject-out cross validation scheme described in section 3.1 was repeated 20 times with the different random initialization of the network weights, resulting in different recognition rates.

TABLE 6
COMPARISON OF RECOGNITION RATE USING SPATIAL FEATURE REPRESENTATION WITH THE CNNS (THE PART 1 IN FIG. 1) LEARNED WITH DIFFERENT OBJECTIVE TERMS IN MMI DATASET AND CASME II DATASET

Objective terms	MMI		CASME II	
	Recognition rates (%)	*p-value*	Recognition rates (%)	*p-value*
E_1	50.87	-	40.65	-
E_1+E_2	65.90	<0.0001	48.37	0.0005
E_1+E_2+E_3	67.63	0.0032	54.07	0.0006
E_1+E_2+E_3+E_4	68.79	0.0071	54.88	0.4909
All (E_1+E_2+E_3+E_4+E_5)	**69.94**	0.0129	**58.54**	0.0100

Note: E_1: minimizing expression classification error, E_2: minimizing intra-class variation, E_3: minimizing expression-state classification error, E_4: minimizing expression-state variation, E_5: preserving expression-state continuity. The p-value indicates the significance of the improvement of recognition rate induced by the added objective term.



(a) Conventional CNN (MMI dataset)

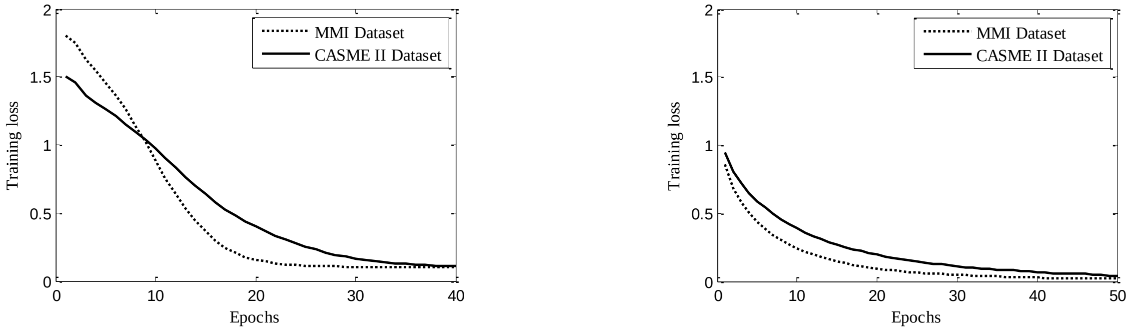
(b) Proposed method (MMI dataset)

(c) Conventional CNN (CASME II dataset)

(d) Proposed method (CASME II dataset)

Fig. 7. A comparative t-SNE visualization [60] of the spatial feature representation extracted by a conventional CNN (a, c) and the proposed spatio-temporal feature representation (b, d). (a) and (b) are the t-SNE visualizations obtained from the deliberate dataset (MMI), (c) and (d) are the t-SNE visualizations obtained from the spontaneous micro-expression dataset (CASME II). Note that, the small dots denote augmented training data and large circles which thick border lines denote testing data. Best viewed in color.

Table 6 shows recognition rates according to different objective terms and the significance of the improvement. In the MMI dataset, the proposed spatial feature representation learning with expression-state constraints improved recognition rate of 19.07% compared to the conventional CNN (using E_1 only). In the CASME II dataset, the recognition rate of the micro-expressions was improved about 17.89 % compared to the conventional CNN. Specifically, utilizing the objective term E_2, which focuses on reducing the intra-class variations of each expression, showed high improvement in the MMI dataset. This is mainly attributed to the characteristics of the expression of MMI dataset, in which there exists a large intra-class variation between different subjects. Furthermore, in recognizing micro-expressions, the results show that the expression-state related objective terms (E_3 and E_5) have a high impact on the recognition rate. The *t*-test results showed that the recognition rate improvements of each objective term were significant in both dataset, except the objective term E_4 in CASME II dataset, which focuses on minimizing expression-state variation. It could be thought that the loss induced by the objective term E_4 in CASME II dataset was small, due to the small variation of frames in the same expression-state class. Note that the recognition rate improvement is significant if the *p-value* is less than 0.05. These results indicate that the proposed objective terms and the frame extraction (selection) of representative expression-state improve classification performance by learning of discriminative spatial feature representation.



(a) (b)

Fig. 8. Training loss curves of the proposed method in MMI dataset and CASME II dataset. (a) CNN in the proposed spatial feature representation learning, (b) LSTM in the proposed temporal feature representation learning.

3.5 Visualizing the Spatio-Temporal Feature Representation Learned with the Proposed Method

To visualize the effectiveness of the proposed spatio-temporal feature representation, t-distributed stochastic neighbor embedding (t-SNE) [60] was utilized. For the comparison, spatial feature representation of the conventional CNN was learned with the apex frames which have strong expression (the conventional CNN is learned by using the objective term E_1). The visualization was performed on both the MMI dataset and the CASME II dataset. In this experiment, for the visualization, subjects of each dataset were randomly divided into two sets; 50% of the training set and 50% of the test set.

Fig. 7 shows the t-SNE visualization of both the spatial feature representation obtained with a conventional CNN and the proposed spatio-temporal feature representation on both the MMI dataset and the CASME II dataset. As shown in the figure, the proposed spatio-temporal feature representation led to more separated classes compared to the spatial feature representation obtained with conventional CNN. These results show that the proposed spatio-temporal representation learning with expression-state constraints produces discriminative feature representation that can be utilized for accurate FER.

3.6 Convergence of the Learning

In this section, the convergence of the proposed learning method was analyzed. Fig. 8 shows the training loss values along the epochs for the CNN and the LSTM used in the proposed method. As can be seen in the figure, the training in MMI dataset converged faster compared to the training in CASME II dataset. This implies facial dynamics in a deliberate expression is easier to learn compared to a spontaneous micro-expression. The result showed that the CNN and the LSTM used in the proposed method converged well on both the MMI dataset and the CASME II dataset.

4 CONCLUSION

In this paper, spatio-temporal feature representation learning for FER that is robust to the expression intensity variations was proposed. For the robust FER, the proposed method utilized representative expression-states in the network trianing which could be specified in facial sequences regardless of the expression intensity variations or different expression durations. In the proposed method, characteristics of facial expressions were encoded in two parts. In the first part, spatial image characteristics of the representative expression-state frames were learned by the CNN. In the first part, four objective terms were proposed to deal with variations in expression classes and expression-states. Moreover, a new objective term was proposed to preserve the expression-state continuity of spatial feature representation for further sequence learning. In the second part, temporal characteristics of the spatial feature representation were consecutively learned with an LSTM. By adopting an LSTM recurrent neural network, the proposed method could learn time scale dependent information in facial expression sequences which have different number of frames.

Extensive and comprehensive experiments have been conducted on the deliberate expression dataset (MMI) and the micro-expression dataset (CASME II). Experiments showed that the proposed method achieved a higher level of FER rates in both datasets compared to the performances with hand-crafted feature based methods as well as existing deep learning based methods.

For the practical FER, the proposed method can be applied to real-world environments in which the expression-state information is not available. As showed in the inter-dataset evaluation in Section 3.3, the proposed method could learn versatile spatio-temporal feature representation of micro-expression and deliberate expression in different datasets. In real-world setting, existing datasets (e.g., MMI and CASME II) which have annotations of expression-states can be used for pre-training the proposed framework. To further improve the recognition performance, LSTM layers in the proposed network could be fine-tuned using real-world datasets which do not have annotations of expression-states. Note that the temporal feature representation learning does not use the expression-state annotations. Thus, the proposed method could be used in real-world environments.

ACKNOWLEDGMENT

This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2015R1A2A2A01005724). Yong Man Ro is the corresponding author.

REFERENCES

[1] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, *et al.*, "The computer expression recognition toolbox (CERT)," *IEEE Int'l Conf. and Workshops on Automatic Face & Gesture Recognition (FG)*, 2011, pp. 298-305.

[2] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, pp. 96-105, 2006.

[3] S. H. Lee, W. J. Baddar, and Y. M. Ro, "Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos," *Pattern Recogn.*, 2016.

[4] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*: Oxford University Press, USA, 1997.

[5] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshop*, 2006, pp. 149-149.

[6] A. Ramirez Rivera, J. Rojas Castillo, and O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, pp. 1740-1752, 2013.

[7] S. S. Jha, P. K. Das, and S. B. Nair, "Automatic facial expression recognition using extended ar-lbp," *Wireless Networks and Computational Intelligence*, ed: Springer, 2012, pp. 244-252.

[8] M.-w. Huang, Z.-w. Wang, and Z.-L. Ying, "A new method for facial expression recognition based on sparse representation plus LBP," *International Congress on Image and Signal Processing (CISP)*, 2010, pp. 1750-1754.

[9] S. Zhang, X. Zhao, and B. Lei, "Robust facial expression recognition via compressive sensing," *Sensors*, vol. 12, pp. 3747-3761, 2012.

[10] W. Zhen and Y. Zilu, "Facial expression recognition based on local phase quantization and sparse representation," *IEEE Int'l Conf. Natural Computation (ICNC)*, 2012, pp. 222-225.

[11] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," *IEEE Int'l Conf. and Workshops on Automatic Face & Gesture Recognition (FG)*, 2011, pp. 314-321.

[12] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 915-928, 2007.

[13] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. Cybernetics*, vol. 44, pp. 161-174, 2014.

[14] S. H. Lee and Y. M. Ro, "Partial Matching of Facial Expression Sequence Using Over-complete Transition Dictionary for Emotion Recognition," *IEEE Trans. Affect. Comput.*, 2015.

[15] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Efficient Spatio-Temporal Local Binary Patterns for Spontaneous Facial Micro-Expression Recognition," *PLoS One*, vol. 10, p. e0124674, 2015.

[16] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 2037-2041, 2006.

[17] C. H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikainen, "(Multiscale) Local Phase Quantisation histogram discriminant analysis with score normalisation for robust face recognition," *IEEE Int'l Conf. Computer Vision Workshops (ICCV Workshops)*, 2009, pp. 633-640.

[18] L. Shen, L. Bai, and M. Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification," *Image and Vision Computing*, vol. 25, pp. 553-563, 2007.

[19] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," *Asian Conf. Computer Vision (ACCV)*, 2014, pp. 143-157.

[20] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l1 optimization," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshop*, 2010, p. 32.

[21] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 974-989, 1999.

[22] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, pp. 172-187, 2007.

[23] J. Chen, D. Chen, Y. Gong, M. Yu, K. Zhang, and L. Wang, "Facial expression recognition using geometric and appearance features," *Proc. IEEE Int'l Conf. Internet Multimedia Computing and Service*, 2012, pp. 29-33.

[24] S. H. Lee, K. Plataniotis, N. Konstantinos, and Y. M. Ro, "Intra-Class Variation Reduction Using Training Expression Images for Sparse Representation Based Facial Expression Recognition," *IEEE Trans. Affect. Comput.*, vol. 5, pp. 340-351, 2014.

[25] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," *Handbook of face recognition*, ed: Springer, 2005, pp. 247-275.

[26] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *IEEE Winter Conf. on Applications of Computer Vision*, 2016, pp. 1-10.

[27] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," *Proc. ACM on Int'l Conf. on Multimodal Interaction*, 2015, pp. 467-474.

[28] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," *Proc. Int'l Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 65-72.

[29] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang, "How Deep Neural Networks Can Improve Emotion Recognition on Video Data," *arXiv preprint arXiv:1602.07377*, 2016.

[30] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," *IEEE Int'l Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1-6.

[31] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired Deep Networks for Facial Expression Feature Learning," *Neurocomputing*, vol. 159, pp. 126-136, 2015.

[32] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition," *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, 2015, pp. 2983-2991.

[33] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," *Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 246-253.

[34] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, *et al.*, "A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2014, pp. 520-527.

[35] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, *et al.*, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS One*, vol. 9, p. e86041, 2014.

[36] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *IEEE Int'l Conf. on Multimedia and Expo (ICME)*, 2005, pp 317-321.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735-1780, 1997.

[38] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recogn.*, vol. 36, pp. 259-275, 2003.

[39] W.-J. Yan, S.-J. Wang, Y.-H. Chen, G. Zhao, and X. Fu, "Quantifying micro-expressions with constraint local model and local binary pattern," *Workshop at the European Conf. on Computer Vision*, 2014, pp. 296-305.

[40] J. Lewis, "Fast normalized cross-correlation," *Vision interface*, 1995, pp. 120-123.

[41] D.-M. Tsai and C.-T. Lin, "Fast normalized cross correlation for defect detection," *Pattern Recogn. Letters,* vol. 24, pp. 2625-2631, 2003.

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278-2324, 1998.

[43] C. M. Bishop, *Neural networks for pattern recognition*. New York: Oxford university press, 1995.

[44] J. Bouvrie, "Notes on convolutional neural networks," 2006.

[45] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: a theoretical and experimental comparison," *INTERSPEECH*, 2013, pp. 1756-1760.

[46] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 2666-2672.

[47] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw.*, vol. 13, pp. 1450-1464, 2002.

[48] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694-4702.

[49] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action Recognition using Visual Attention," *NIPS workshop on Time Series*, 2015.

[50] Y.-l. Tian, "Evaluation of face resolution for expression analysis," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2004, pp. 82-82.

[51] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1859-1866.

[52] S. Y. Park, S. H. Lee, and Y. M. Ro, "Subtle Facial Expression Recognition Using Adaptive Magnification of Discriminative Facial Motion," *Proc. 23rd Annual ACM Conf. on Multimedia*, 2015, pp. 911-914.

[53] P. Khorrami, T. Paine, and T. Huang, "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?," *IEEE Int'l Conf. on Computer Vision Workshops*, 2015, pp. 19-27.

[54] S. Demyanov. *ConvNet*. Available: https://github.com/sdemyanov/ConvNet

[55] F. Chollet. (2015). *Keras*. Available: https://github.com/fchollet/keras

[56] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proc. Int'l Conf. on Machine Learning*, 2010, pp. 807-814.

[57] Y.-H. Oh, A. C. Le Ngo, J. See, S.-T. Liong, R. C.-W. Phan, and H.-C. Ling, "Monogenic Riesz wavelet representation for micro-expression recognition," *IEEE Int'l Conf. on Digital Signal Processing (DSP)*, 2015, pp. 1237-1241.

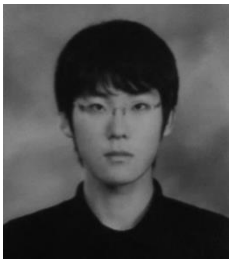
[58] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320-3328.

[59] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *IEEE Int'l Conf. artificial intelligence and statistics*, 2010, pp. 249-256.

[60] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, p. 85, 2008.

Dae Hoe Kim received the B.S. degree from Hanyang University, Seoul, Korea in 2010 and the M.S. and Ph.D degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2017, respectively. In 2012, he was a visiting researcher at the University of Toronto, Toronto, ON, Canada. His research interests include medical imaging, facial expression recognition, pattern recognition, machine learning, and computer vision.

Wissam J. Baddar received the B.S. degree from the University of Jordan, Amman, Jordan in 2010, and the M.S. degree from the KAIST, Daejeon, South Korea, in 2014. He is currently working towards the Ph.D. degree in the Image and Video Systems Laboratory at KAIST. His research interests include face recognition/detection, face expression recognition/analysis, biometrics, medical imaging, pattern recognition, and machine learning.



Jinhyeok Jang received the B.S. degree in 2014 and the M.S. degree in 2016 from Ulsan National Institute of Science and Technology, Ulsan, South Korea. Then, he is working on the Ph.D course from the KAIST, Daejeon, South Korea. His present research interests include image processing, blur estimation, image recognition, deep learning and medical image processing.

Yong Man Ro (S'85-M'92-SM'98) received the BS degree from the Yonsei University, Seoul, Korea, and the MS and PhD degrees from the KAIST, Daejeon, Korea. He was a researcher at the Columbia University, a visiting researcher at the University of California, Irvine, CA, and a research fellow at the University of California, Berkeley, CA. He was a visiting professor in the Department of Electrical and Computer Engineering at the University of Toronto, Canada. He is currently a professor and the chair of signals and systems group of the Department of Electrical Engineering, KAIST. He established Image and video systems (IVY) Lab, KAIST, in 1997. Among the years, he has been conducting research in a wide spectrum of image and video systems research topics. Among those topics; image processing, computer vision, visual recognition, medical image processing, and video representation/compression. He received the Young Investigator finalist Award of ISMRM in 1992, and the year's scientist award (Korea), in 2003. He served as an associate editor for the IEEE Signal Processing Letters. He serves as an associate editor in transitions on data hiding and multimedia security (Springer-Verlag). He served as a TPC in many international conferences including the program chair and organized special sessions. He is a senior member of the IEEE.