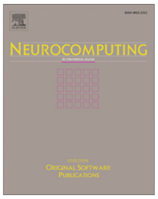
Neurocomputing 549 (2023) 126439

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Deep emotion change detection via facial expression analysis

ByungOk Han [a], Cheol-Hwan Yoo [a], Ho-Won Kim [a], Jang-Hee Yoo [a], Jinhyeok Jang [a],*

[a] ETRI, Republic of Korea

ARTICLE INFO

Article history:
Received 18 October 2022
Revised 16 March 2023
Accepted 5 June 2023
Available online 9 June 2023
Communicated by Zidong Wang

Keyword:
Emotion change detection
Facial expression
Social communication
Temporal spotting

ABSTRACT

Facial expressions are one of the most essential channels to communicate a person's emotional state. In social interaction, the capability to accurately read subtle changes in facial expressions, which reveal emotional fluctuations, is critical for 1) comprehending others' emotions in context and background situations, 2) identifying responsiveness to others' emotions, and 3) developing social skills in human–computer interaction. In this paper, we first introduce automatic emotion change detection via facial expression that discovers timings or temporal locations in a video where facial expression significantly changes. We propose a weakly-supervised deep emotion change detection framework that does not require facial expression videos with expensive temporal annotations and instead learns static images for training. Incorporating these ideas, we performed extensive experiments to demonstrate fundamental insights into emotion change detection and the efficacy of our framework using three video datasets, i.e., CASME II, MMI, and our YoutubeECD. Furthermore, we modified our framework for temporal spotting, which is the most similar task to emotion change detection, and showed comparable results with state-of-the-art methods on CAS(ME)^2, proving justification for the problem. Even though we only employed the AffectNet to train our framework rather than the CASME II, MMI, YoutubeECD, and CAS(ME)^2, experimental results demonstrate its exceptional generalization capability in cross-dataset environments.

1. Introduction

Non-verbal communication, i.e., non-language-mediated communication, involves the use of non-verbal cues such as facial expressions, eye contact, gestures, posture, tone of voice, and body language. Non-verbal communication is vital in human interactions because it can serve as a guide, an amplifier, or even a substitute for verbal communication [4–6]. In general, non-verbal and verbal communication complement each other, but non-verbal aspects in face-to-face interactions may express a completely different meaning [7]. For example, saying "stop" with a smile or a neutral facial expression may convey an entirely different intention. Facial expressions play a significant role in social interactions among those non-verbal cues because they provide visual cues for sending and perceiving emotional states via human faces [8].

From this point of view, for nearly three decades, automatic facial expression analysis has been extensively studied in a variety of fields because of its practical importance in social robots, smart video surveillance, patient monitoring, driver monitoring, self-management interviews, and other human–computer interaction (HCI) systems [9]. Particularly, research on automatic facial expression recognition has primarily focused on what facial expressions are being made. That is, a user's emotional state is deduced by an automatic emotion recognition system based on the facial expressions he or she makes. For instance, [10]'s method discriminates between emotions, i.e., happiness, sadness, disgust, anger, surprise, fear, and neutrality, by learning and generating the corresponding neutral face image for any facial expression image. With the success of deep learning technology, numerous emotion classification algorithms based on facial expression analysis have been developed, as described in the survey paper [11]. Meanwhile, [12] proposed a multi-modal emotion recognition system that uses facial expressions, shoulder gestures, and audio cues in valence-arousal space based on the dimensional emotion model [13].

To properly understand human emotions through facial expressions, not only is the automatic emotional state recognition task crucial for non-verbal communication, but also is automatic emotion change detection task for three fundamental reasons:

1. *The timing and direction, i.e., positive or negative shift, of emotion change are important factors in understanding emotion status appropriately with contextual and background situations* [14–16].

* Corresponding author.
E-mail addresses: byungok.han@etri.re.kr (B. Han), ch.yoo@etri.re.kr (C.-H. Yoo), hw_kim@etri.re.kr (H.-W. Kim), jhy@etri.re.kr (J.-H. Yoo), jjh6297@etri.re.kr (J. Jang).

https://doi.org/10.1016/j.neucom.2023.126439
0925-2312/© 2023 The Author(s). Published by Elsevier B.V.

For example, if a person is hospitalized and has been experiencing persistent sad feelings, the sadness is amplified when he or she receives unexpected negative news about his or her health. In this case, automatic emotion recognition merely detects the user's emotional state, i.e. sadness. That is, automatic emotion recognition alone misses important contextual and background information, which has a significant impact on the user's emotional state.

2. *The timing of emotion change provides critical clues for identifying responsiveness to others' emotions in social communication* [17]. Responding to others' emotions is a complex skill that heavily relies on social and cognitive abilities. People with mental disorders such as Autism Spectrum Disorder (ASD) and alexithymia, who have a poor ability to understand others' emotions, are known to be unable to respond to others' emotions automatically and immediately [18,19]. For instance, during ASD screening, automatic emotion change detection in social interaction can provide an objective indicator of emotional response.

3. *It is critical for natural HCI systems with social capabilities to detect when the user's emotional state changes.* Non-verbal means of communication, i.e., contextual information such as facial expression, intonation, behavior, gaze, and gesture, should be interpreted at a certain point in time in order to fully comprehend the user's emotional state. That is, the timing of the user's emotional change can be considered as a critical time for grasping the user's emotional state, which should be comprehended along with the context and background situations. For sociable HCI, results of automatic emotion change detection can be used as reference points, i.e., beginning or changing points, of a computer's interaction method with humans.

Automatic emotion change detection via facial expression finds timings or temporal locations in a video where facial expressions shift significantly, as shown in Fig. 1. Despite the importance mentioned above, few studies have focused on the automatic emotion change detection task. [20] reported an initial investigation into detecting and localizing changepoints based on speech signals using the Gaussian Mixture Model (GMM). In [21], the automatic emotion change detection method for the speech modality is also proposed using a martingale framework. However, to the best of our knowledge, no research on the automatic emotion change detection task via facial expressions has been conducted except for our workshop paper [22], which we extend to this paper.

In the fields of computer vision and pattern recognition, the closest study to emotion change detection via facial expressions is the temporal spotting in facial expression videos, which is also known as temporal localization or segmentation. The temporal spotting problem is similar to the emotion change detection problem in that the goal is to find key parts of facial expression videos. However, the fundamental difference is that the emotion change detection finds changepoints, whereas the temporal spotting finds intervals where facial expressions occur over time. In general, the temporal dynamics of facial expressions are modeled by onset (i.e., starting point), apex (i.e., peak point), and offset (i.e., ending point) [23] over time. Temporal spotting identifies intervals based on the facial expression temporal model. On the other hand, the emotion change detection does not assume the temporal model and only identifies changepoints of facial expressions. The temporal model of facial expressions does not cover complex and subtle changes in internal emotions due to the following practical reasons [24]:

1. *The temporal model cannot account for all of the temporal dynamics of spontaneous expressions, e.g., onset-onset-offset and onset-offset-offset, which occur frequently in a complex manner in real environments.* For example, if a user laughs a little and then laughs a lot or cries a little and then cries a lot, the simple onset-offset model cannot handle it, resulting in missing important context information. In CAS(ME)^2 [25], for instance, 44 intervals out of 374 facial expressions have no offset label, indicating that annotators were not able to seek them out. Fig. 2 is a counterexample of the temporal model in CAS(ME)^2, where its offset label is missing.

2. *In real-world situations, the entire temporal sequence, i.e. neutral-onset-apex-offset-neutral, may not always be available* [26]. Thus, Acted Facial Expressions in the Wild (AFEW) dataset does not assume the full temporal dynamics of facial expressions for close-to-natural settings. In addition, even if there is an offset in a micro-expression, it may not appear in the video due to the camera's frame-rate limitation.

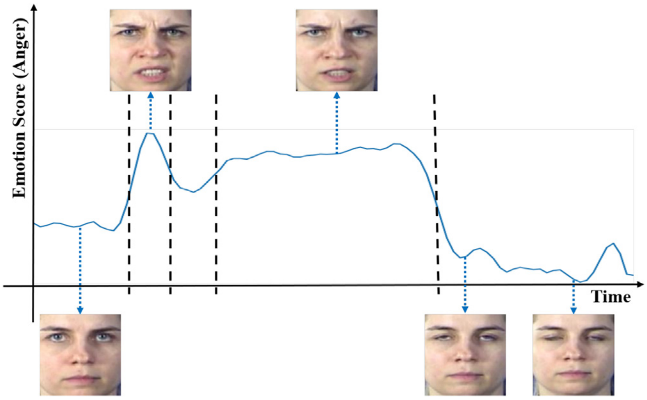


Fig. 1. Detection of emotion changes by analyzing an emotion signal from a facial expression video on MMI dataset [3]. The emotion change detection task identifies timings or locations in a video where facial expressions significantly shift. Estimated changepoints provide crucial meaning for comprehending human emotions. The emotion signal describes the intensity changes of participant's anger emotion from our framework. The dotted lines represent changepoints detected in the signal.

As a result, from a facial expression video, the onset and offset points obtained by an automatic spotting algorithm are a subset of the changepoints obtained by an automatic emotion change detection method. To demonstrate the justification for the problem, we modified our framework to produce proper results in the spotting task by utilizing the relationship between the two tasks, as described in Section 3.6. Moreover, we conducted extensive comparative experiments with state-of-the-art algorithms for the spotting task of the Micro Expression Grand Challenge (MEGC) as described in Section 4.6.

In this paper, we present, to the best of our knowledge, the first study to investigate the automatic emotion change detection (ECD) problem via facial expressions from an engineering perspective. To tackle the issue, we propose a weakly-supervised Deep Emotion Change Detection (DECD) framework comprised primarily of a multi-task emotion recognizer and a changepoint detector. The Multi-Task Emotion Recognizer (MTER) learns discrimination of categorical facial expressions as well as valence and arousal representations. The MTER extracts an emotion signal, which consists of frame-level emotion states, using recent Deep Neural Networks (DNNs) from a facial expression video. Then, our changepoint detector identifies points with relatively steep and large changes based on the emotion signal over time. We provide a comprehensive framework for online and offline ECD tasks that includes a variety of recent DNNs and several existing changepoint detection approaches.

In summary, the main contributions of our work are:

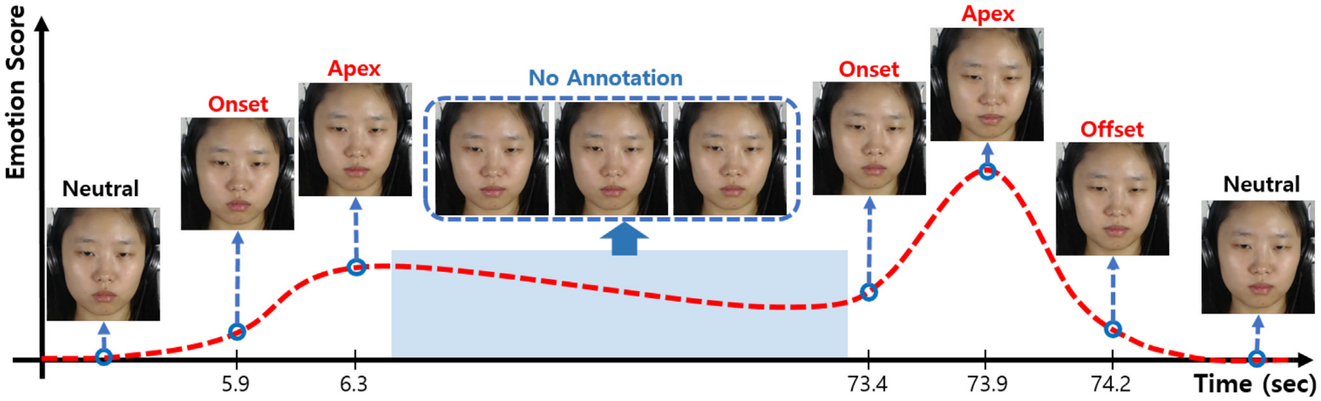


Fig. 2. A counterexample of facial expression temporal model in CAS(ME)^2 dataset. In this video, there are two macro-facial expressions with very slight motions around the participant's mouth, but only the second one has an offset label. The temporal model cannot account for all of the temporal dynamics of spontaneous expressions in real environments. Our proposed framework for emotion change detection is able to identify the changepoints regardless of the temporal model of facial expressions.

- We first cast the facial expression-based automatic ECD problem. We provide a formal definition, practical justification, and correlation with the most similar task, i.e., temporal spotting, for the problem.
- We propose a weakly-supervised DECD framework that only learns static facial expression images rather than directly learning video clips in which emotions change. This strengthens our method because there are fewer temporal labels, which require a laborious annotation process, than facial expression labels.
- We demonstrate extensive experimental results that provide fundamental insights into the facial expression-based ECD task and validate our ideas on a static image dataset for training, i.e., AffectNet [1], and three video datasets for testing, i.e., CASME II [2], MMI [3], and our own YoutubeECD.
- We adapted our DECD framework and conducted comparative experiments with state-of-the-art methods for the temporal spotting task. Without any data from CAS(ME)^2-cropped for training, we show comparable results for the spotting task of MEGC.

The remainder of this paper is organized as follows. Section 2 introduces the related work, including temporal spotting for facial expression, changepoint detection, and deep facial expression recognition. In Section 3, we formulate the facial expression-based ECD problem and then describe our comprehensive framework and application to the temporal spotting problem. Our extensive experimental results, including the performance of our DECD framework and comparisons with recent temporal spotting methods, are presented in Section 4. Then, we conclude and discuss our work in Section 5.

2. Related Work

2.1. Temporal Spotting for Facial Expression

Temporal spotting for facial expressions is the task of identifying onset-offset intervals in a video. The spotting task is the most similar to our ECD task, but there are fundamental differences as mentioned in Section 1. Recently, the temporal spotting task for the MEGC has attracted many researchers using various DNN based data-driven approaches. Pan et al. [27] applied a bilinear Convolutional Neural Network (CNN) which is composed of two parallel streams for feature extraction and a Support Vector Machine (SVM) as a classifier. Wang et al. [28] used a spatio-temporal CNN that consists of a 2D-CNN for extracting frame-wise feature vectors from a video and a 1D-CNN for mixing the feature vectors for the temporal spotting task. Yang et al. [29] also adopted a facial action unit-based deep learning framework for the spotting task. Those methods train their DNNs on CAS(ME)^2 and SAMM [30] datasets in a supervised manner. However, due to the lack of datasets, those fully-supervised approaches struggle to collect facial expression video samples and their temporal annotations, i.e., onset, peak, and offset frames, for training DNNs. This limits the generalization capability in a cross-dataset experiment setting. Our DECD framework, on the other hand, is trained in a weakly-supervised manner, using only static facial expression images and their emotion labels rather than facial expression videos and their temporal labels.

2.2. Changepoint Detection

Changepoint detection has been researched to identify steeply changing points in a sequential signal along the temporal dimension since the 1950s [31,32]. Changepoint detection is divided into two categories: offline and online detection. Offline changepoint detection algorithms discover changepoints after the whole signal has been obtained. There are numerous widely used offline methods, such as binary segmentation [33], sliding window [34], fused Lasso [35], bottom up [36], and Pruned Exact Linear Time (PELT) [37].

Online changepoint detection aims to find changepoints in real-time scenarios. Different from offline approaches, online ones should determine if the current point is a changepoint or not by using only signal information from the past to the present. Online approaches, in particular, use less data than offline approaches to detect changepoints in streaming signals, making it more difficult to make correct decisions. Early online approaches were developed under the assumption that the distribution of signal data was known [31,32]. In [38,39], online detection has been generalized by automatic analysis of the distribution of signal data. Based on Bayesian theory, an online method of [40] was proposed, where it recursively estimates the probabilistic distribution to determine whether a changepoint occurs or not. The Bayesian approach is being actively modified to improve its performance and is being used in various areas [41,42]. In our DECD framework, we deal with representative changepoint methods with qualitative and quantitative experimental results for offline and online ECD tasks.

2.3. Deep Facial Expression Recognition

Facial Expression Recognition (FER) has received steady attention for several decades. In particular, with the development of

DNNs, CNN-based approaches based on static facial expression images, e.g., [43–45], have been intensively investigated for facial expression classification based on six basic emotions [46]. In addition, video-based FER approaches are proposed using CNN [47,48], Recurrent Neural Network (RNN) [49], and Long Short-Term Memory (LSTM) [50]. Nasir et al. [51] utilized the fuzzy membership to capture gradual change in human emotion for emotion classification in facial videos. Recently, with promising results from Vision Transformer (ViT) [52] for image recognition, the transformer-based DNN structures began to be applied to FER [53,54]. We train the MTER of our DECD framework using several representative DNNs and validate it for the facial expression-based ECD task.

The amount and quality of FER data used for training DNNs have a significant impact on the performance of DNN-based FER methods. Early FER datasets [55,3] are composed of controlled and posed facial expression image or video data collected in laboratory environments, whereas more recent FER datasets [56,1] are composed of unrestricted data collected in real environments. The unrestricted FER datasets, also known as in-the-wild datasets, include various facial variations in identity, pose, illumination, and occlusion because they collect diverse facial expression images or videos by searching with emotion keywords on the Internet. Using AffectNet [1], the largest in-the-wild FER dataset, we train our MTER to learn the FER classification and the dimensional information on valence-arousal space for properly depicting emotional states. Furthermore, we constructed our own in-the-wild YoutubeECD dataset to validate our DECD framework in real environments.

3. Proposed Method

In this section, we begin by formulating the ECD problem, including the problem's goal and performance criteria. Our DECD framework consists of three parts, a face detector, an MTER, and a changepoint detector, as shown in Fig. 3. Face detection is performed from a video input containing facial expression information. Based on detected regions, the MTER estimates an emotional state per frame, which consists of emotion scores, an arousal value, and a valence value. The frame-level emotion states extracted from the MTER are concatenated to generate an emotion signal per video, which is then used for emotion change analysis. Finally, the emotion signal is analyzed to detect multiple changepoints in a multivariate time series by changepoint detection. The details are given below.

3.1. Problem Definition

Emotion changes that occur within a person or are caused by another person provide important clues for understanding a person's feelings. As mentioned in Section 1, the timings of emotion changes can be utilized as key indicators for determining a person's emotional intelligence or sociable HCI. *The goal of the facial expression-based ECD problem is to discover or locate those timings or points in time from a video containing human faces where facial expression shifts significantly or steeply.* Based on the facts, we evaluate how accurately an automatic ECD approach predicts the timing of emotion changes. To be specific, we define ground truth changepoints $\mathbf{T}_{gt}$ and predicted changepoints $\mathbf{T}_{pred}$ in a video of a person's facial expressions as:

$$\begin{aligned}\mathbf{T}_{gt} &= \{t_1, t_2, \ldots, t_{n_{gt}}\}\\ \mathbf{T}_{pred} &= \{t_1, t_2, \ldots, t_{n_{pred}}\}\end{aligned} \tag{1}$$

where n_{gt} is the number of ground truth changepoints in a video and n_{pred} is the number of predicted changepoints from an automatic ECD method.

Each ground truth changepoint is compared to all of the predicted changepoints as shown in Fig. 4 and then precision and recall of the ECD method are defined as:

$$\begin{aligned}Prec. &= \sum_{i=1}^{n_{gt}} min\left(1, \sum_{j=1}^{n_{pred}} tp(t_i, t_j)\right)/n_{gt}\\ Recall &= \sum_{i=1}^{n_{gt}} min\left(1, \sum_{j=1}^{n_{pred}} tp(t_i, t_j)\right)/n_{pred}\\ tp(t_i, t_j) &= \begin{cases} 1 & \text{if } |t_i - t_j| \leqslant \tau \\ 0 & \text{otherwise}\end{cases}\end{aligned} \tag{2}$$

where, $tp(\cdot)$ is a function that determines true positive and τ is a time interval threshold in seconds. τ should be set to a short value that reflects the time between changepoints. Note that a duplicated true positive sample is protected by the function $min(\cdot)$.

We employed the threshold τ without using the error between the ground truth and predicted value for the evaluation criterion. This is because whether or not a person's emotional changes can be detected within an appropriate time, i.e., τ, is an important indicator in social communication. We use precision rather than recall as a quantitative metric in Sections 4.4 and 4.5. This is because we fix the number of predictions to a predefined number for easy-to-understand analysis for the ECD task.

3.2. Multi-task Emotion Recognition

For several decades, psychologists have developed diverse emotion models to find out how human emotions work. We adopted a categorical model [46] and a dimensional model [13] among them because of their simplicity and generality. The categorical model is a typical emotion model that represents universal facial expressions based on Paul Ekman's emotion theory. Seven discrete emotion categories are widely used, which include happiness, sadness, disgust, fear, anger, surprise, and neutrality. On the other hand, the dimensional model was developed by Russell and represents an emotional state in terms of valence and arousal dimensions rather than in discrete categories. The two emotion models describe different aspects of an emotional state in humans and can play a complementary role in each other.

We designed our deep learning architecture in a multi-task manner, which covers categorical and dimensional models, to represent a complex emotional state of humans for each video frame. Multi-task learning techniques have several advantages compared to single-task learning methods. One of the most significant benefits is that associated tasks complement one another to improve the representations of features in input vectors. The categorical model can represent each emotion by discriminating several emotions in an independent direction, but it cannot represent the magnitude of each emotion. The dimensional one, on the other hand, is available to represent the magnitude of each emotion, but only has two directions: valence and arousal. In other words, the categorical model is expected to improve inter-class discriminability while the dimensional model is expected to clarify intra-class magnitude ordering. In addition, multi-task learning improves computational efficiency and generalization ability in feature space. Taking those advantages, our method performs emotion state recognition in a multi-task manner based on the two emotion models as follows.

Given input facial images $\mathbf{x}$ and their labels $\mathbf{y}_e, \mathbf{y}_v$, and $\mathbf{y}_a$, we constitute a feature extractor $f(\mathbf{x}; \theta_f)$ with a trainable parameters θ_f. Then, an emotion classifier is defined as $e(f(\mathbf{x}); \theta_e)$, consisting of the output of the feature extractor with a trainable parameter, θ_e. Similarly, a valence regressor is represented as $v(f(\mathbf{x}); \theta_v)$ and an arousal regressor is formulated as $a(f(\mathbf{x}); \theta_a)$. The loss function $E(\cdot)$ of our MTER, $MTER(\mathbf{x}, \mathbf{y}_e, \mathbf{y}_v, \mathbf{y}_a; \theta_f, \theta_e, \theta_v, \theta_a)$, is defined as:

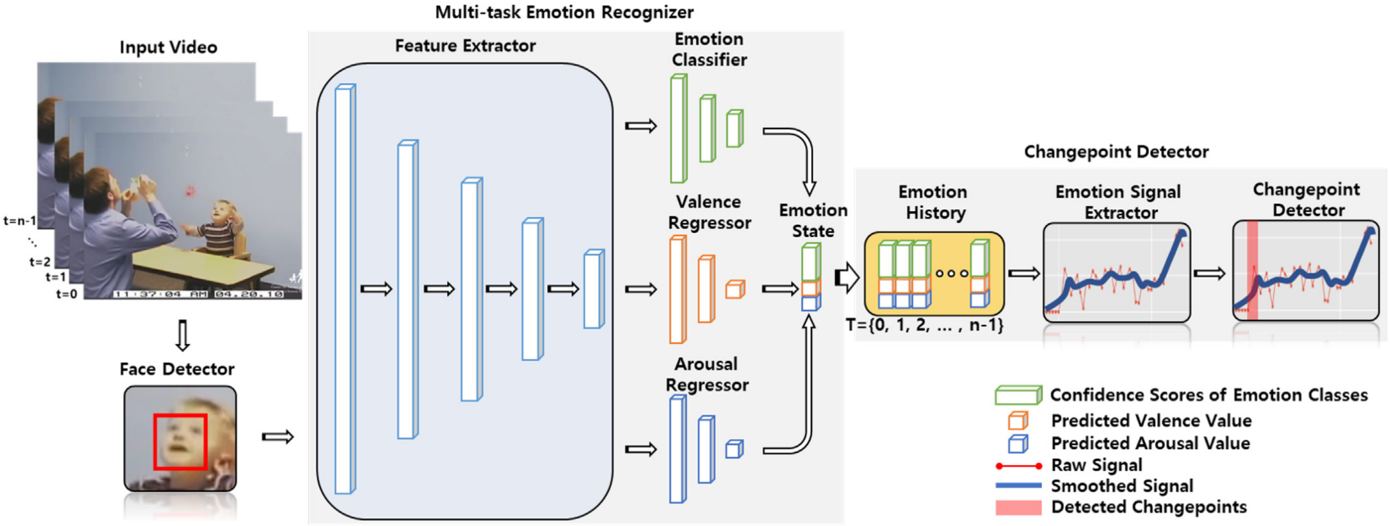


Fig. 3. The overall architecture of our framework. Our DECD framework consists of a face detector, a multi-task emotion recognizer (MTER), and a changepoint detector. The MTER produces an emotion signal, which is composed of emotion scores and valence-arousal values, and the emotion signal is analyzed by the changepoint detector to extract multiple changepoints per video sample.

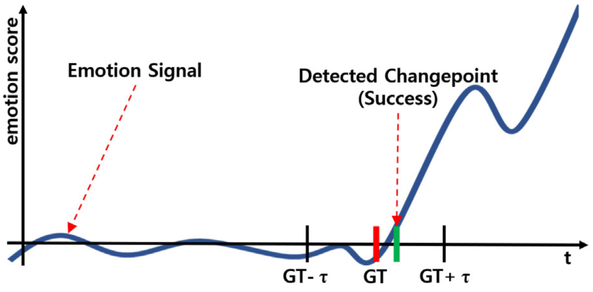


Fig. 4. Evaluation criterion of emotion change detection problem. GT represents the ground truth of a changepoint and τ is a time interval threshold. Each ground truth changepoint is compared to all of the detected changepoints based on the threshold τ.

$$\begin{aligned} &E(MTER(\mathbf{x}, \mathbf{y}_e, \mathbf{y}_v, \mathbf{y}_a; \theta_f, \theta_e, \theta_v, \theta_a)) \\ &\quad = \lambda_e CCE(e(f(\mathbf{x})), \mathbf{y}_e; \theta_f, \theta_e)) + \lambda_v L_2(v(f(\mathbf{x})), \mathbf{y}_v; \theta_f, \theta_v) \\ &\qquad + \lambda_a L_2(a(f(\mathbf{x})), \mathbf{y}_a; \theta_f, \theta_a) \end{aligned} \tag{3}$$

where, $\mathbf{y}_e, \mathbf{y}_v$, and $\mathbf{y}_a$ stand for labels of the emotions, valence, and arousal of facial images $\mathbf{x}$. $CCE(\cdot)$ depicts the loss function of categorical cross entropy for the multi-class classification problem. $L_2(\cdot)$ describes a function of mean squared error and λ_e, λ_v, and λ_a are weight values for each task.

The emotional state deduced from the MTER is represented as a nine-dimensional vector, which consists of seven dimensions for confidence values of categorical emotions, one dimension for a valence value, and one dimension for an arousal value. After estimating the nine-dimensional emotion state, the frame-level emotion state is accumulated into an emotion signal that illustrates emotional information from video input.

3.3. Weakly Supervised Learning

To properly train DNNs, it is necessary to prepare large-scale datasets with accurate annotations. However, annotating temporal labels on a large amount of facial expression video data is a time-consuming process. In addition, there exists substantial ambiguity in annotating facial expression temporal labels, which leads to inconsistency in temporal labeling. For these reasons, facial expression video data with temporal labels is quite insufficient to appropriately train the MTER for the ECD task. To overcome these problems, we train our DECD framework in a weakly-supervised manner using static facial expression images with categorical and dimensional emotion annotations, which are much more sufficient than video data with temporal labels.

Weakly supervised learning is a way to learn a high-level task, such as ECD on facial expression videos, from a low-level task, such as emotion recognition on static facial expression images. To be specific, the MTER in our DECD framework uses 2D DNN architectures, instead of 3D DNNs, as our backbones that can easily learn generic visual features through pre-training from large-scale image datasets, such as ImageNet [57]. The visual features from large-scale datasets should help generalization performance of our framework on facial expression videos for the ECD task. For this, we initialize the MTER using a pre-trained model on ImageNet. Based on the initialized MTER, instead of directly learning facial expression videos with their temporal labels, our DECD framework learns static facial expression images with their emotion state labels for frame-level video encoding, resulting in its remarkable generalization capability in cross-dataset environments as described in Section 4.

3.4. Noise Filtering

Once an emotion state signal is obtained, the DECD framework performs a noise filtering process to remove outliers, such as high-frequency noise, and identify the general tendency of emotional flow from the signal using a Savitzky-Golay filter [58]. Since our framework determines changepoints based on the emotion score retrieved from each frame in a video, it can result in high-frequency noise. To be specific, it may be caused by the instability of the trained model or by the prediction noise from variations in illumination, facial pose, and partial occlusions of facial images. The Savitzky-Golay filter is a sort of low-pass filter well-known for its characteristics of reducing noise while maintaining the height of the signal shape and peaks of a waveform. Thus, the smoothing filter is necessary to reduce inessential details in the signal. The discrete nine-dimensional vector of the emotion state signal with n frames in a video can be defined as:

$$\mathbf{s}_i = \{s_i(0), s_i(1), \ldots, s_i(n-1)\} \quad \text{s.t. } 0 \leqslant i < 9 \tag{4}$$

The filter performs a moving polynomial fit on $2M+1$ points. The polynomial function $p(t)$ with order N is denoted as:

$$p(t) = \sum_{k=0}^{N} w_k t^k \tag{5}$$

where, w_k depicts a coefficient of the polynomial function $p(t)$.

Then, local least-square approximation error ε_N between the signal and the polynomial function can be described as:

$$\varepsilon_N = \sum_{t=-M}^{M} (p(t) - s_i(t))^2 \tag{6}$$

A filtered signal r_i is calculated by using a discrete convolution operation as:

$$\begin{aligned} r_i(t) &= \sum_{m=-M}^{M} h[m] s_i(t-m) \\ &= \sum_{m=t-M}^{t+M} h[t-m] s_i(m) \end{aligned} \tag{7}$$

where $h[\cdot]$ denotes a finite impulse response that is equivalent to the least-square polynomial approximation.

The filtered signal $\mathbf{r}_i$ is obtained by repeatedly performing the convolution operation with a moving window of $2M+1$ width on the emotion state signal $\mathbf{s}_i$.

3.5. Changepoint Detection

After the noise filtering process, our method conducts changepoint detection on the emotion signal. Changepoint detection algorithms are categorized into offline or online methods depending on whether they are performed on a whole signal from entire frames or on a partial signal in real-time. When analyzing recorded video inputs, we employ an offline method to detect changepoints based on entire video frames. On the other hand, when detecting emotional changes in a real-time situation such as a robot, we utilize an online method to obtain immediate results.

3.5.1. Offline changepoint detection

We use a bottom-up merge approach for offline changepoint detection because of its simplicity and effectiveness. The bottom-up merge method splits a whole signal into a large number of short sub-signals. Then, it repeatedly merges consecutive sub-signals if the cost difference between the sub-signals is comparatively small, which implies there is no large emotion changes. Those processes continue until the specified number of changepoints is reached or the cost difference is greater than a threshold value. Fig. 5 shows how the bottom-up merge approach works for offline changepoint detection.

To be specific, we first measure the amount of change in the signal to find points where the emotion signal changes. The degree of change in the signal, or homogeneity, is measured to quantify whether there are drastic changes in the signal itself. To reflect the degree of changes in the nine-dimensional emotion signal $\mathbf{r}$, we integrate it into a one-dimensional signal r^* by applying the median filter to it along the dimension axis. Using the radial basis function (RBF) [59], the cost function $C(\cdot)$ for a homogeneity measure of a sub-signal $r^*(t_s, t_e)$ from t_s to t_e is defined as:

$$C(r^*(t_s, t_e)) = (t_e - t_s) - \frac{1}{t_e - t_s} \sum_{j=t_s+1}^{t_e} \sum_{k=t_s+1}^{t_e} exp\left(-\gamma \|r^*(j) - r^*(k)\|^2\right) \quad \text{s.t. } \gamma > 0 \tag{8}$$

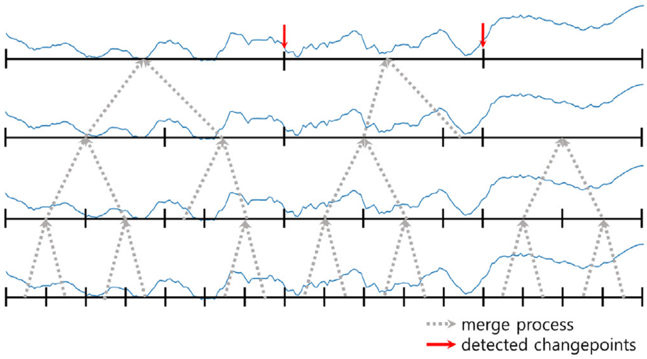


Fig. 5. Bottom-up merge approach for offline changepoint detection. It repeatedly merge consecutive sub-signals until a stopping criteria is satisfied. In this case, the stopping criteria is that the number of changepoints is two points.

where, γ denotes a bandwidth parameter for the radial basis kernel.

The cost function produces a low value if a signal contains few changepoints. A large cost value means that a signal has several changepoints. Using the cost function, the distance between two consecutive sub-signals can be obtained as:

$$\text{Dist}(r^*(t_s, t_m), r^*(t_m, t_e)) = C(r^*(t_s, t_e)) - C(r^*(t_s, t_m)) - C(r^*(t_m, t_e)) \quad \text{s.t. } 0 \leqslant t_s < t_m < t_e \leqslant n-1 \tag{9}$$

The distance between consecutive sub-signals increases when the cost of an entire signal is high and the total cost of its subsequent sub-signals is low, indicating that the two signals are heterogeneous. After obtaining all the distances between successive sub-signals, the two signals with the smallest distance are combined repeatedly. If the number of changepoints is known and set by a user, our method stops the iterative merge process when the number of sub-signals reaches the number of changepoints. Otherwise, the stopping criteria based on the cost difference between the two sub-signals is defined as:

$$\text{Dist}(r^*(t_s, t_m), r^*(t_m, t_e)) \geqslant pen \tag{10}$$

If the penalty value, *pen*, is small, our method sensitively responds to changes in the signal, resulting in many detected changepoints. Conversely, if the *pen* is large, our method tries to detect a few significant changepoints.

3.5.2. Online changepoint detection

We also developed an online changepoint detection method to find emotional changes based on the Bayesian theory [40] for real-time applications. The bayesian online changepoint detection approach identifies anomaly changes based on history information at regular intervals. When a new emotion state value $r_i(t+1)$ is received, a posterior probability is calculated based on a previous signal $\mathbf{r}_i(0, t)$ as follows.

$$\begin{aligned} P(r_i(t+1) \mid \mathbf{r}_i(0,t)) &= \sum_{\ell_t} P(r_i(t+1), \ell_t \mid \mathbf{r}_i(0,t)) \\ &= \sum_{\ell_t} P(r_i(t+1) \mid \ell_t, \mathbf{r}_i(0,t)) P(\ell_t \mid \mathbf{r}_i(0,t)) \end{aligned} \tag{11}$$

where ℓ_t denotes the run lengths, indicating time length at t since the last changepoint, for calculating a marginal probability over it.

The predictive probability $P(r_i(t+1) \mid \ell_t, \mathbf{r}_i(0,t))$ can be solved by using a recursive form as in Eq. (13). Then, we have the posterior probability $P(\ell_t \mid \mathbf{r}_i(0,t))$ to be solved.

$$P(\ell_t \mid \mathbf{r}_i(0,t)) = \frac{P(\ell_t, \mathbf{r}_i(0,t))}{P(\mathbf{r}_i(0,t))} \tag{12}$$

The posterior is proportional to the joint probability $P(\ell_t, \mathbf{r}_i(0,t))$ and it can be derived to the recursive form as follows.

$$\begin{aligned} P(\ell_t, \mathbf{r}_i(0,t)) &= \sum_{\ell_{t-1}} P(\ell_t, \ell_{t-1}, \mathbf{r}_i(0,t-1), r_i(t)) \\ &= \sum_{\ell_{t-1}} P(\ell_t, r_i(t) \mid \ell_{t-1}, \mathbf{r}_i(0,t-1)) P(\ell_{t-1}, \mathbf{r}_i(0,t-1)) \\ &= \sum_{\ell_{t-1}} P(\ell_t \mid \ell_{t-1}) P(r_i(t) \mid \ell_{t-1}, \mathbf{r}_i(0,t-1)) \\ &\quad \times P(\ell_{t-1}, \mathbf{r}_i(0,t-1)) \end{aligned} \tag{13}$$

When a changepoint is detected, the run length ℓ_t is reset to 0, otherwise it becomes $\ell_{t-1}+1$. Then, the conditional prior in terms of the run length $P(\ell_t \mid \ell_{t-1})$ can be obtained by using the hazard function $H(\cdot)$ [60].

$$P(\ell_t \mid \ell_{t-1}) = \begin{cases} H(\ell_{t-1}+1) & \text{if } \ell_t = 0 \\ 1 - H(\ell_{t-1}+1) & \text{if } \ell_t = \ell_{t-1}+1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where we use a constant hazard function $H = 1/\lambda$ that is not dependent on the run length ℓ_t and the time t. A user parameter λ denotes an expected length between two changepoints in the emotion signal.

To calculate the posterior predictive in Eq. (13), we assume that the sub-signal follows a normal likelihood with an unknown mean and variance as:

$$(r_i(t) \mid \ell_{t-1}, \mathbf{r}_i(0,t-1)) \sim \mathcal{N}(\mu, \sigma^2) \tag{15}$$

Using this assumption, we can calculate the posterior predictive as:

$$\begin{aligned} &P(r_i(t) \mid \ell_{t-1}, \mathbf{r}_i(0,t-1)) \\ &\quad = \int_{\mu,\sigma^2} \mathcal{N}(r_i(t) \mid \mu, \sigma^2) P(\mu, \sigma^2 \mid \ell_{t-1}, \mathbf{r}_i(0,t-1)) \mathrm{d}(\mu, \sigma^2) \end{aligned} \tag{16}$$

However, this integral is difficult to compute, so we use the conjugate before getting a derived closed-form [61]. The distribution of the conjugate prior in terms of the normal distribution is the Normal-Gamma distribution. Then, the posterior predictive can be obtained through the Student-T distribution [62]:

$$P(r_i(t) \mid \ell_{t-1}, \mathbf{r}_i(0,t-1)) = t_{2\alpha}\left(r_i(t) \mid \hat{\mu}, \frac{\beta(K+1)}{\alpha K}\right) \tag{17}$$

where $\alpha, \beta, \hat{\mu}, K$ are the posterior parameters. The posterior predictive follows a Student-T distribution with a center at $\hat{\mu}$, precision $\alpha K/\beta(K+1)$, and degree of freedom 2α.

When a new emotion state value is received, the four parameters in Eq. (17) are updated every time as:

$$\begin{aligned} &K \quad \to K+1, \alpha \quad \to \alpha + 0.5, \\ &\hat{\mu} \to \tfrac{K\mu + x}{K+1}, \quad \beta \quad \to \beta + \tfrac{K}{K+1} \tfrac{(x-\hat{\mu})^2}{2} \end{aligned} \tag{18}$$

Finally, using the Student-T distribution, we can obtain the posterior probability for change prediction, i.e., $P(r_i(t+1) \mid \mathbf{r}_i(0,t))$. We perform change prediction by comparing the $(t+1)$-th posterior and the t-th posterior along the temporal dimension. For the nine-dimensional emotion signal, we average over the difference between the two posterior probabilities as follows:

$$\frac{1}{n_d} \sum_{i=0}^{n_d-1} (P(r_i(t+1) \mid \mathbf{r}_i(0,t)) - P(r_i(t) \mid \mathbf{r}_i(0,t-1))) < \psi \tag{19}$$

where, n_d is the number of dimensions of the emotion signal and ψ is a threshold value for the bayesian decision.

The Bayesian method treats the point as a changepoint in the emotion signal if this criterion is satisfied.

3.6. Application to Temporal Spotting

Our DECD framework can be applied to the facial expression spotting task, which identifies intervals where macro and micro facial expressions occur. We consider the intervals obtained by facial expression spotting, which is composed of onset and offset points, to be a subset of detected emotion changepoints from our framework. Most facial expression studies assume that emotional changes occur only in the temporal model of facial expressions, which are composed of onset, peak, and offset. Some changepoints, however, are not described by the temporal model of facial expressions [26]. For example, if a facial expression occurs at a very brief moment in a video sample, such as a micro-expression, only the onset is visible and the offset is not. The CAS(ME)^2 [25] contains tens of samples that lack offset labels because annotators discovered the onset and peak points but not offset points. Taking these factors into account, we consider the detected points from the ECD task to be a superset of those from the facial expression spotting task.

For the facial expression spotting task, we need to remove unnecessary changepoints to detect onset-offset intervals. A set of the detected emotion changepoints from our framework $\mathbf{T}_{ecd}$ is defined as:

$$\mathbf{T}_{ecd} = \{t_1, t_2, \ldots, t_{n_p}\} \tag{20}$$

where, n_p represents the number of detected changepoints.

Then, all intervals related to facial expression dynamics $\mathbf{I_{ecd}}$ are represented as a combination of the detected changepoints $\mathbf{T}_{ecd}$:

$$\begin{aligned} &\mathbf{I}_{ecd} = \{(t_1, t_2)_1, (t_1, t_3)_2, \ldots, (t_{n_p-1}, t_{n_p})_{n_i}\} \\ &n_i = C^2_{n_p} \end{aligned} \tag{21}$$

where $C^2_{n_p}$ is the number of all intervals by a combination of the detected changepoints $\mathbf{T}_{ecd}$.

To find intervals composed of onset, peak, and offset, we filter out intervals that do not follow the temporal dynamics of facial expressions. A peak point t_{peak} of an interval in $\mathbf{I}_{ecd}$ is obtained by using an onset point t_{on} and an offset point t_{off} as:

$$\begin{aligned} &t_{peak} = floor\left(\tfrac{t_{on}+t_{off}}{2}\right) \\ &\text{s.t. } 0 \leqslant t_{on} < t_{peak} < t_{off} < n_p \end{aligned} \tag{22}$$

where, *floor* denotes a floor function to obtain an integer value.

We can obtain a set of spotted intervals $\mathbf{I}_{spot}$ from the set $\mathbf{I}_{ecd}$ by selecting only the intervals that satisfy the following condition:

$$\frac{|min(\mathbf{r}_i(t_{on}) - \mathbf{r}_i(t_{peak}), \mathbf{r}_i(t_{off}) - \mathbf{r}_i(t_{peak}))|}{t_{off} - t_{on}} \geqslant \eta \tag{23}$$

where, $\mathbf{r}_i$ is one dimension of the nine-dimensional emotion signal used as a reference signal to calculate peak points and emotion state values. η denotes a threshold value for removing intervals that do not follow the temporal dynamics of facial expressions.

In this way, our DECD framework can be used to produce the set of spotted intervals $\mathbf{I}_{spot}$, which consists of onset and offset pairs in a facial expression video. This modified version of our DECD framework, DECD-spot, produces appropriate results for the temporal spotting task, as described in Section 4.6, by exploiting the relationship between the ECD and temporal spotting tasks. Not only that, but those experimental results demonstrate the justification of the ECD problem.

4. Experimental Results

Generally, the temporal model of emotion changes in facial expression is modeled by onset (i.e., starting point), apex (i.e., peak point), and offset (i.e., ending point) [23] along the temporal dimension. Based on the temporal dynamics of facial expression, we assume that changes in emotions occur at onset, apex, and offset points according to facial movements. This is because facial expression datasets, which we used for experiments such as CASME II and CAS(ME)^2, only provide temporal labels for onset, apex, and offset indices, not changepoint labels. Furthermore, MMI dataset provides videos of posed facial expressions in laboratory environments, so participants for MMI made their facial expressions according to the temporal model. However, our YoutubeECD dataset does not assume the temporal model of facial expressions and provides only a changepoint per video clip. In order to clarify the ECD task, we set the goal of detecting emotion change to finding an onset frame in a facial expression video in this section except for Section 4.6. We consider a changepoint in a YoutubeECD video as an changepoint in it.

In this section, all of the ECD and temporal spotting experiments are carried out in cross-dataset settings. For validation of our DECD framework, we tested it on CASME II, MMI, CAS(ME)^2, and our own YoutubeECD datasets without training on split sets from them. *To be specific, even though their video data and annotations were created in totally different environments, we did not train our DECD framework on CASME II, MMI, CAS(ME)^2, or our own YoutubeECD datasets; instead, we only used AffectNet. This is a significant advantage of our weakly-supervised DECD framework since it demonstrates exceptional generalization capacity of our framework for the ECD problem.* The description of the datasets as well as the implementation details, ablation study, main results, and comparison experiments with the other approaches are given below.

4.1. Datasets

For evaluation of the ECD task, we utilized five datasets, i.e., AffectNet [1], CASME II [2], MMI [3], YoutubeECD, and CAS(ME)^2 datasets. To train the MTER, we used the AffectNet dataset by using 7-class emotion categories and values for valence and arousal. To evaluate the overall DECD framework, we used CASME II, MMI, and YoutubeECD datasets, which have ground truths of onset or changepoint along the temporal dimension. For comparative experiments with state-of-the-art methods, we employed the CAS(ME)^2 [25] dataset for the facial expression spotting task. The CASME II, MMI, and CAS(ME)^2 datasets are recorded in lab-controlled environments, where facial expressions in videos are captured by several fixed cameras. On the other hand, our YoutubeECD dataset provides online videos that have spontaneous and natural facial expressions in unrestricted environments from the Youtube website. Details of the datasets are described as follows.

4.1.1. AffectNet

AffectNet is an in-the-wild facial expression dataset that is collected by querying 1,250 emotion-related keywords on search engines such as Google, Bing, and Yahoo. AffectNet is the largest facial expression dataset which contains more than 1 M facial images from the Internet and annotations of facial expressions, valence, and arousal. We used a total of 287,401 facial images, which have annotations for seven classes (i.e., neutrality, happiness, sadness, surprise, disgust, fear, and anger) out of eleven emotion categories and values for valence and arousal. AffectNet officially provides training and validation sets which have 283,901 and 3,500 facial images, respectively. Using this dataset, we train our multi-task recognizer, which includes a seven-class classifier for facial expressions and a regressor for valence and arousal.

4.1.2. CASME II

CASME II is a micro-expression video dataset recorded at the high temporal resolution, 200 fps, in the lab environment. The micro-expressions are very rapid and subtle facial expressions that usually last between 0.04 and 0.5 secs [63–65] while the macro-expressions occur between 0.5 and 4 s. There are a total of 255 video clips which consist of five classes of micro-expressions, i.e., surprise, repression, happiness, disgust, and others. This dataset officially provides the temporal dynamics information, which includes the frame number of onset, apex, and offset along the temporal dimension in image sequences. We utilized all of the video clips in the dataset with cropped faces and annotations of the onset frame to evaluate our method.

4.1.3. MMI

MMI is an in-the-lab facial expression dataset that contains over 2,900 image sequences recorded at 25 fps from 75 subjects. To test ECD approaches, we used 127 image sequences related to five emotion categories (i.e., happiness, sadness, disgust, fear, and anger), 30 subjects, low-resolution images, and frontal faces. All of the image sequences in this dataset are manually and tightly segmented from the beginning (i.e., onset frame) to the end (i.e., offset frame) of the facial expressions. However, for testing ECD approaches, we need full videos, which must contain all the temporal dynamics from the neutral state of facial expressions to the onset, apex, and offset, with their annotations for temporal dynamics such as the frame number of the onset in each video. To evaluate detection of the onset point in a video clip, we reconstructed the video by replicating the first frame image, which has a neutral face, 15 times at the beginning of it. Then, we annotated the 16th frame as the onset frame, which is used for the ground truth information in the ECD task.

4.1.4. YoutubeECD

To evaluate our method in-the-wild environment, we constructed the YoutubeECD dataset, which contains emotional changes of ASD children or infants. This dataset is composed for analysis of social interactions and non-verbal communications related to facial emotion changes in ASD children or infants. In general, children or infants with ASD have difficulty in emotional interaction with others since the way and timing of revealing emotions tends to be different from those with TD [18]. In this regard, detecting emotion changes can provide essential clues for ASD screening of children or infants.

We collected a total of 461 videos by querying keywords, such as autism, autism spectrum disorder, child, and infant, on the Youtube website. Since the YoutubeECD dataset is composed of facial videos, which have emotion changes, from unrestricted environments, they inherently have large variations in facial expression, identity, pose, illumination, and occlusion as shown in Fig. 6. This is a very important property as a test environment for developing ECD algorithms that are robust to various variations. We also annotated the frame number of a changepoint per facial expression video clip, three classes of emotional categories (positive, negative, and neutral), and expression intensity ranging from 0 to 3. All video clips are segmented with a margin of around 1.0 to 3.0 secs based on the changepoint we annotated. For testing ECD methods, we used 134 video clips that have annotations of positive and negative facial expressions. Due to the privacy concerns when using

Fig. 6. Examples of the YoutubeECD. We blurred faces in captured images of videos due to the privacy issue.

facial images of children and infants, we did not open our YoutubeECD dataset to the public.

4.1.5. CAS(ME)^2

CAS(ME)^2 [25] is a dataset for the spotting task of spontaneous macro-expression and micro-expression. The dataset contains 87 long videos recorded at 30 fps with annotations of onset, peak, and offset indices along the temporal dimension. We employ this dataset to compare our proposed method to the-state-of the-art methods for the facial expression spotting task. To be specific, we use the CAS(ME)^2-cropped version, which officially provides face regions of long videos from Micro-Expression Grand Challenge (MEGC) 2021 [66]. The CAS(ME)^2-cropped dataset includes 300 macro-expressions and 57 micro-expressions with an average duration of 148 s.

4.2. Implementation Details

Based on the fact that the micro-expression occurs for a short moment ($\leqslant 0.5$ secs) [65], we set the τ to 0.2 secs on the evaluation criterion, Eq. (2) for Sections 4.4 and 4.5. This ensures that a detected changepoint does not deviate from the range of micro-expression. Our method first detects facial regions in a video input using RetinaFace [67]. All detected facial images are resized to 100×100. For multi-task emotion recognition, we utilized a ResNet [68] network as our backbone network. Once the nine-dimensional emotion state vectors are extracted from the MTER as described in Section 3.2, we concatenate the vectors to compose an emotion signal, which will be analyzed to find multiple changepoints. To train the MTER, we initialized our MTER using a pre-trained model on ImageNet and fine-tuned it on AffectNet using the stochastic gradient descent optimizer with a batch size of 512 and the number of training epochs of 100. The initial value of the learning rate was set to 0.01 and it decreases by 10 times every 30 epochs.

After composing the emotion signal, we conducted a noise filtering process to remove outliers in the signal through the Savitzky-Golay filter [58] as described in Section 3.4. We used 17 frames as the width of the moving window and a 13-degree polynomial to fit the extracted signal to the filter.

For changepoint detection, we employed bottom-up, binary segmentation, sliding window, dynamic programming, and PELT approaches to quantitatively compare results for offline changepoint detection. We implemented those algorithms by using a python library, "Ruptures" [69]. For online approaches, we implemented the Bayesian approach and a gradient-based approach to predict whether an emotional change occurs or not in real-time. For the Bayesian online changepoint detection method, we used a library, "Baysian Online Changepoint Detection in Python" [40], and initialized the posterior parameters to $\mu = 0, K = 1, \alpha = 1, \beta = 1$. The ψ is set to zero for the Bayesian decision.

For an easy-to-understand analysis, we performed experiments based on the number of changepoints detected by those changepoint detection methods. In more detail, in the cases of binary segmentation, bottom-up, sliding window, dynamic programming, and gradient-based methods, we conducted experiments assuming that the number of changepoints is known. That is, those five methods find as many changepoints as the number of changepoints specified. Note that, among the offline methods, the binary segmentation, bottom-up, and sliding window approaches can perform change detection by setting either the *pen* value or the number of changepoints as described in Section 3.5.1. On the other hand, because the PELT method is not based on the assumption, the number of detected changepoints was controlled by adjusting the *pen* value. The Bayesian method was tested based on the λ value because the number of changepoints detected is more dependent on the λ value than β value as described in Section 3.5.2.

For the temporal spotting task, we used ResNet-18 as an MTER backbone, bottom-up with a *pen* as a changepoint detector, and RBF as a cost function without the noise filtering process. As a reference signal of Eq. (23), we empirically set the 'Happiness' signal from the nine-dimensional emotion signal.

4.3. Temporal Dynamics of Emotion Changes in Facial Expression

We first performed qualitative experiments for the visualization of signal outputs from our DECD framework. Our method extracts an emotion signal from a video clip that has faces and their facial movements, and analyzes it to detect emotional changepoints, which are start points or breakpoints that separate sub-signals along the temporal dimension of the signal. Fig. 7 shows results from our emotion changepoint detection method using two video samples in MMI and CASME II datasets. Note that the samples from MMI and CASME II are recorded at 25 fps and 200 fps respectively. Due to the recording speed, extracted emotion signals have differences in the temporal resolution of graph visualization.

The moments of emotional changes in the upward or downward direction appear consistently along the temporal dimension in all of the signals. In particular, in the case of the anger signal from the MMI sample, the value of the signal changes steeply upward at the moment of onset. Similarly, signals from the other emotional dimensions also show steep changes at the onset point. Based on each multivariate signal, our method successfully detected an onset point (red line) within τ (green shading) in Eq. (2), i.e., 0.2 secs, compared to one of the ground truth (green line).

4.4. Ablation Study

We conducted experiments on the ablation study to validate improved performance by changing or removing key components of our framework. To be specific, we verified the effectiveness of our backbone network for the MTER, the noise filtering, the cost function, and the signal type of the emotion signal for our DECD framework. In this ablation study, in order to clarify the ECD task, we set the goal of detecting emotion change to find an onset frame in a facial expression video. We used CASME II dataset because there is no public ECD dataset. A video sample on CASME II has one facial expression occurrence per sample with the ground truth information of an onset in facial expression dynamics. We tested our method on onset detection based on the assumption that the number of changepoints is known and that there are two signifi-

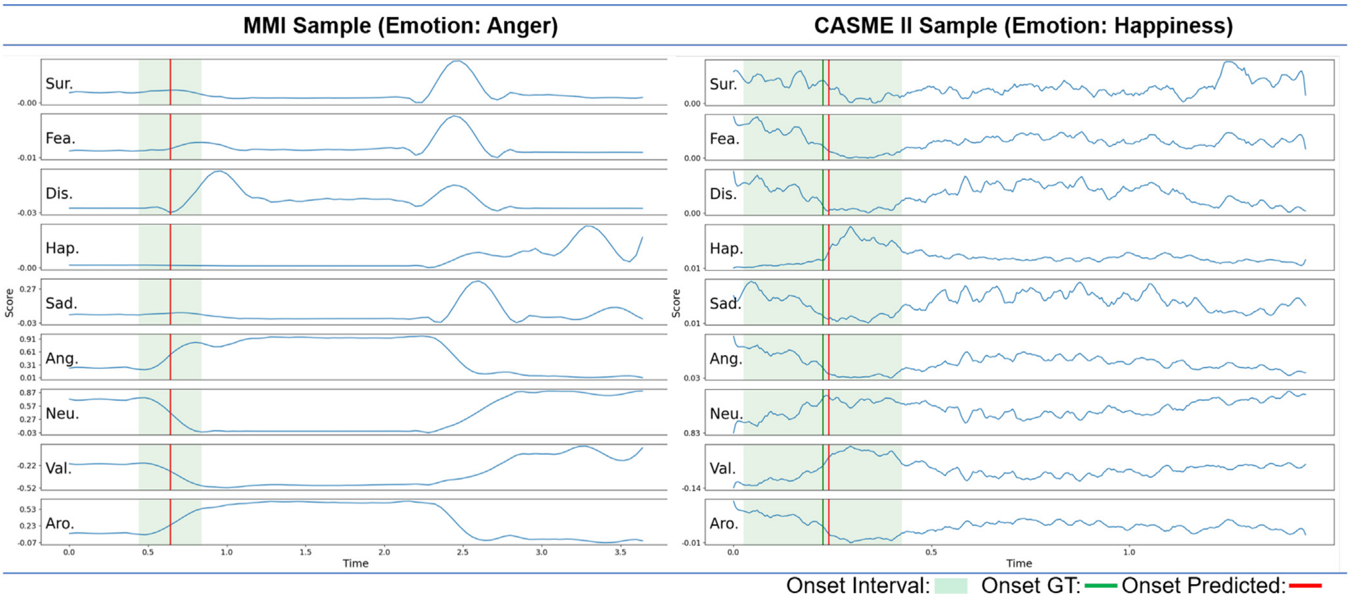


Fig. 7. Emotion signals extracted from two video samples, one annotated as an angry expression from the MMI dataset and one with happiness from the CASME II dataset. The emotion signal consists of nine dimensions of seven emotion classes, valence, and arousal. The green line represents the ground truth (GT) of the onset point in facial expression dynamics, and the red line denotes a detected point from an offline changepoint detection method. The green shading indicates the ground truth interval for the onset point. Each row shows a signal from one out of the nine dimensions of an emotion signal. (Sur.: Surprise, Fea.: Fear, Dis. :, Disgust, Hap.: Happiness, Sad.: Sadness, Neu: Neutrality, Val.: Valence, Aro.: Arousal).

cant changepoints, i.e. onset and offset points, per sample on CASME II. Namely, we set the number of changepoints to two, which is composed of onset and offset points along the temporal dimension, and the first point detected by our method is used for an estimated onset point as described in Fig. 7. Because the quantity of predictions from our algorithm is set, precision is utilized to assess its performance rather than recall, which is also fixed.

4.4.1. DNNs for multi-task emotion recognition

The MTER provides a nine-dimensional emotion signal from a facial video. This is achieved by training our backbone DNN using the train set of the facial images in the AffectNet dataset with labels for seven emotion classes and dimensional values for valence and arousal. We verified the quality of a frame-level emotion state, which composes an emotion signal, on the validation set of AffectNet and the test set of RAF [56] by training several DNNs widely used for facial expression recognition. We used the RAF test set to evaluate cross-dataset performance of the MTER when trained on the AffectNet train set. We also conducted the onset detection experiments on CASME II dataset to evaluate ECD performance. As can be seen in Table 1, SwinT shows the highest performance in the seven-class emotion classification and regression of valence and arousal on the AffectNet. However, the Transformer-based architectures such as ViT and SwinT do not perform well on the RAF test set and CASME II due to a lack of inductive bias. ResNet-50 achieves the highest precision in the onset detection task because of its exceptional generalization ability in a cross-dataset setting. Note that the precision does not increase as the error decreases. This is because the ECD task aims to predict within an appropriate time, as mentioned in Section 3.1. We used ResNet-50 as our backbone for the MTER in the other ablation experiments.

4.4.2. Noise filtering

Since each emotional state is extracted at the frame level, an emotion signal extracted from the MTER would be noisy. The signal noise is caused by the identity of facial images, changes in facial pose, illumination changes, occlusion, and other factors. To reduce signal noise while preserving the signal's global shape, we employ noise filtering methods. We investigated the impact of three representative filtering algorithms that smooth a signal on the onset detection task. The Savitzky-Golay filter achieves the highest precision with the largest gap from the raw signal as shown in Table 2.

4.4.3. Cost function for distance measure

We also conducted the ECD experiments to validate the cost functions in a changepoint detection algorithm. A cost function of changepoint algorithms measures the homogeneity of sub-signals to find changepoints in an emotion signal. The homogeneity is utilized to distance measure between sub-signals in an emotion signal. We employed six cost functions to verify the performance of our method, as shown in Table 3. Among them, the L1-median cost function produces the best precision on the ECD task. The L1-median cost function is generally robust to a shift in the median value of a sub-signal [76]. In other words, because steep changes in emotion signals occur frequently, the L1-median that reflects these changes in the cost value produces good results. However, the RBF shows the best performance in average error, but not in ECD precision.

4.4.4. Emotion signal type

In addition, we performed experiments to verify the effectiveness of the extracted nine-dimensional emotion signal. We composed six types of emotion signals as 'Happiness', 'Neutrality', 'Emotion', 'Predicted Emotion', 'Valence & Arousal', and 'All' based on the extracted signal as shown in Table 4. The two signal types of 'Neutrality' and 'Emotion' show comparatively high performance. Because the values of the 'Neutrality' signal type are converted to probability values using the softmax function at the final layer of the MTER, they are closely related to the remaining six emotion signal values in the 'Emotion' one. In other words, the 'Neutrality' signal type can be regarded as a representative signal of the 'Emotion' signal type. The 'Predicted Emotion' signal type is composed of maximum values along the temporal dimension in one of 'Emo-

Table 1
Ablation study on various DNN backbones to extract a frame-level emotion state on onset detection. The results show the accuracy of the emotion classification task and mean squared errors of regression tasks for valence and arousal on the validation set of AffectNet and the test set of RAF

Backbone Networks	Emo. Acc. (%)		Val. MSE	Aro. MSE	ECD Prec. (%)	Avg. Err. (sec)
Dataset	Aff.	RAF	Aff.	Aff.	CASME II	CASME II
RegnetX [70]	58.3	66.7	0.098	0.110	81.96	**0.1236**
ViT-base [71]	58.0	50.0	0.094	0.099	78.43	0.1555
SwinT [72]	**59.9**	50.0	**0.092**	**0.094**	80.78	0.1438
MLP-Mixer [73]	53.1	**83.3**	0.117	0.115	79.61	0.1405
ResNet-18 [74]	58.7	75.0	0.097	0.107	83.92	0.1302
ResNet-50	57.6	75.0	0.104	0.100	**84.71**	0.1291
ResNet-152	57.3	**83.3**	0.107	0.108	71.76	0.1823

(Emo.: Emotion, Acc.: Accuracy, Val.: Valence, Aro.: Arousal, MSE: Mean Squared Error, Prec.: Precision, Err.: Error, Aff.: AffectNet)

Table 2
Ablation study on noise filtering for onset detection on CASME II. Note that we assume the number of changepoints in a video is known and is fixed at 2. The first detected changepoint is used for onset detection.

Filtering Methods	Kernel Size	Poly-order	σ	ECD Prec. (%)	Avg.Err. (sec)
Raw Signal	-	-	-	82.75	0.1278
Gaussian	17	-	1	83.14	0.1272
Bilateral [75]	17	-	1	83.92	**0.1269**
Savitzky-Golay [58]	17	13	-	**84.71**	0.1291

Table 3
Ablation study on cost function for onset detection on CASME II. Note that we assume the number of changepoints in a video is known and is fixed at 2. The first detected changepoint is used for onset detection.

Cost Functions	ECD Prec. (%)	Avg. Err. (sec)
Normal	79.61	0.1346
Cosine	83.14	0.1317
Linear	78.43	0.1572
RBF [59]	82.75	**0.1275**
L2-mean	83.92	0.1287
L1-median	**84.71**	0.1291

Table 4
Ablation study on emotion signal type for onset detection on CASME II. Dim. denotes the dimension of a signal. Note that we assume the number of changepoints in a video is known and is fixed at 2. The first detected changepoint is used for onset detection.

Signal Type	Dim.	ECD Prec. (%)	Avg. Err (sec)
Happiness	1	72.55	0.1840
Neutrality	1	82.35	0.1444
Emotion	7	83.14	0.1319
Predicted Emotion	1	79.22	0.1578
Valence & Arousal	2	81.96	0.1373
All	9	**84.71**	**0.1291**

tion'. Among the signal types, the 'All' signal type achieves the best performance, which indicates that all the nine-dimensional information in an emotion signal is effectively applied to the ECD task.

4.5. Performance of Changepoint Detection Algorithms

In this subsection, we report the main experimental results to validate the efficacy of our DECD framework in various environments using two in-the-lab datasets, i.e., CASME II and MMI, and an in-the-wild dataset, i.e., our YoutubeECD dataset. In detail, based on our proposed framework, we verify the ECD performance of several offline and online changepoint detection approaches on hundreds of video clips with macro and micro facial expressions. Similarly to Section 4.4, we perform onset detection experiments here, but we control the user's threshold, such as *pen* or the number of changepoints in Section 3.5.1 so that the number of detected changepoints ranges from 1 to 10, rather than limiting it to 2. In other words, along the temporal dimension, all detected changepoints are compared to the onset point's ground truth in a video clip.

On the extracted emotion signals, we test binary segmentation (BinSeg) [33], sliding window (Window) [34], dynamic programming (DynP) [77], pruned exact linear time (PELT) [37], and bottom-up (BottomUp) [36] methods for offline changepoint detection. We also perform experiments on simple gradient-based and Bayesian approaches for online changepoint detection. To fairly compare the performance of those methods, we modified online methods into offline methods. In the case of the gradient-based method, a changepoint was not detected by simply comparing a gradient value of the emotion signal with a user's threshold, but the changepoint was detected based on the ranking of gradient values in the entire signal. The Bayesian approach was evaluated based on the λ of the constant hazard function H in Eq. (14), a user parameter for the expected length, to show performance more directly.

Despite the fact that almost all the CASME II video samples contain micro-expressions, which last in 0.5 secs, the overall algorithms show relatively high performance on it, as shown in Table 5. This is because video samples on the CASME II are recorded at 200 fps, resulting in high temporal resolution of extracted emotion signals. Among changepoint algorithms, the DynP approach produces the highest precision since it yields the optimal solution to discrete optimization [69]. However, the DynP guarantees the optimal solution only when the number of changepoints is known. That is, in most cases where the ECD is used, this assumption is incorrect. Among approaches that do not require the assumption, the BottomUp method shows the best performance. In online changepoint methods, the Bayesian approach performs better than the Gradient method when comparing performance when the average number of detected changepoints is one.

Table 6 shows the performance of the online and offline methods on the MMI dataset. Since samples on the MMI dataset are recorded at 25 fps and the temporal resolution is low, the overall performance of algorithms on it is lower than on the CASME II. The DynP and BottomUp methods also show higher precision than the other methods in the onset detection task of macro expressions. As shown in Section 4.1.3, the ground truth of the onset in

Table 5
Onset detection performance on CASME II dataset. Note that the Avg. Prec. represents average precision calculated without columns that include missing values. Avg. # DCPs describes the average number of detected changepoints. The normalized precision, Norm. Prec., is computed by dividing the Prec. by the Avg. # DCPs which represents the precision per prediction.

Algorithms			Avg. # DCPs										Avg.
			1	2	3	4	5	6	7	8	9	10	Prec.
Offline	BinSeg		50.59	85.88	**91.76**	94.12	95.69	97.65	98.43	99.22	99.22	-	89.17
	Window		46.27	66.67	79.22	86.33	89.22	93.93	96.04	97.43	-	-	81.89
	DynP		50.59	**87.45**	**91.76**	**95.29**	**97.65**	98.43	98.82	**100**	**100**	-	**90.00**
	PELT		38.04	63.67	75.08	84.04	91.37	92.82	94.43	95.52	96.75	97.78	79.37
	BottomUp		**52.16**	86.27	91.37	94.12	**97.65**	**98.82**	**99.61**	99.61	99.61	**100**	89.95
Online	Gradient		36.47	54.51	67.84	75.69	81.96	84.71	88.63	89.80	90.59	91.76	72.45
	Bayesian	λ	3	5	7	10	13	16	19	22	25	30	-
		Avg. # DCPs	1.00	2.01	2.64	2.63	2.61	3.47	14.33	10.35	20.84	21.64	-
		Prec.	44.31	48.24	55.69	60.78	67.45	73.33	95.29	94.90	92.55	81.57	-
		Norm. Prec.	**44.31**	24.00	21.10	23.11	25.84	21.13	6.65	9.17	4.44	3.77	-

Table 6
Onset detection performance on MMI dataset.

Algorithms			Avg. # DCPs										Avg.
			1	2	3	4	5	6	7	8	9	10	Prec.
Offline	BinSeg		**36.22**	62.99	73.23	78.74	83.46	88.19	90.55	-	-	-	71.65
	Window		26.77	55.12	72.00	80.56	-	-	-	-	-	-	69.23
	DynP		**36.22**	62.99	**74.02**	**81.89**	85.83	**89.76**	**91.34**	-	-	-	**72.97**
	PELT		-	57.74	70.76	79.00	84.23	88.28	90.26	92.78	95.37	96.61	69.17
	BottomUp		35.43	**64.57**	73.23	80.31	**86.61**	88.19	88.19	89.76	92.13	92.91	72.70
Online	Gradient		0	2.36	13.39	29.92	48.03	69.29	82.68	**95.28**	**96.85**	**98.43**	15.22
	Bayesian	λ	3	5	7	10	13	16	19	22	25	30	-
		Avg. # DCPs	1.00	2.17	2.77	2.93	3.48	4.99	11.95	8.06	9.81	8.53	-
		Prec.	0	0	74.02	100	0	0	0	0	0	0	-
		Norm. Prec.	0	0	26.72	**34.13**	0	0	0	0	0	0	-

the MMI dataset is fixed at 16th frame, so the Bayesian method performs best when the expected length λ of the sub-signal is appropriate. In general, this is not a reliable result because the duration of macro and micro expressions is not fixed and varies from 0.04 secs to 4 s. However, when the expected length of sub-signals is approximately correct, the Bayesian approach produces outstanding performance.

We also evaluate our DECD framework on our in-the-wild dataset, YoutubeECD. In Table 7, the performance of seven approaches on the YoutubeECD is lower than those on the other in-the-lab datasets. Because our framework is based on facial information, the various variations of facial information have a significant impact on performance. Unlike the in-the-lab datasets, the video clips in the YoutubeECD were recorded in unrestricted environments, so the emotion signals are extracted based on facial images reflecting diverse and large variation factors, such as changes in facial identity, facial pose, illumination, occlusions, and so on. Nevertheless, the Window method shows the highest performance among online and offline algorithms. The Window method, on the other hand, is heavily dependent on the length of the sliding window algorithm, which is a user parameter. Therefore, it is ineffective for detecting emotion changes in videos with varying facial expression durations along the temporal dimension. Even in an unrestricted setting, DynP maintains a high level of performance among seven methods.

We also analyze the execution time for changepoint detection methods as shown in Table 8. Note that the execution time does not include time for the multi-task emotion recognition and noise filtering processes. The Gradient and Window methods are faster than the others because of their algorithmic simplicity. High complexity methods, such as the Bayesian, DynP, and PELT methods, on the other hand, are slower than the others.

4.6. Comparisons with the Other Approaches

Finally, we conducted comparative experiments with the other approaches similar to our framework. Because, to the best of our knowledge, our DECD framework is the first attempt to detect emotional changes using facial expression information, there is no comparative study for the task. We conduct comparative experiments by adapting and modifying our framework for the facial expression spotting task, which is most similar to the ECD task as described in Section 3.6. We utilize the CAS(ME)^2-cropped dataset from the MEGC 2021[1] challenge for the spotting task. We use the Intersection over Union (IoU) score to evaluate spotting algorithms quantitatively, where true positive (TP) per interval is defined based on the predicted interval and the ground-truth interval as follows [66]:

$$\frac{I_{pred} \cap I_{gt}}{I_{pred} \cup I_{gt}} \geqslant 0.5 \tag{24}$$

where I_{gt} denotes the ground truth of the facial expression interval between onset and offset points, and I_{pred} represents the predicted interval from a spotting method.

Note that each ground truth interval corresponds to at most one single predicted interval. If Eq. (24) is not satisfied, the sample is classified as a false positive (FP). Then, for recall, precision, and the f1-score metric, we count the number of TP, FP, and false negative (FN) samples for the spotting task.

The nine-dimensional emotion signal from a video sample on the CAS(ME)^2-cropped dataset and the results of spotted intervals from our framework are depicted in Fig. 8. The emotion signal was

[1] MEGC 2021 Homepage:https://megc2021.github.io/

Table 7
ECD performance on YoutubeECD dataset.

	Algorithms		Avg. # DCPs										Avg.
			1	2	3	4	5	6	7	8	9	10	Prec.
Offline	BinSeg		8.96	17.91	20.90	29.10	33.58	44.78	52.24	58.96	64.18	67.16	36.73
	Window		5.22	15.67	22.39	31.46	**43.52**	**51.46**	**60.35**	**67.91**	**75.62**	-	**41.51**
	DynP		8.96	**19.40**	22.39	30.60	39.55	47.01	54.48	58.96	64.93	**69.40**	38.48
	PELT		**10.45**	15.75	24.78	**31.86**	36.23	43.39	50.54	57.78	62.87	64.56	37.07
	BottomUp		6.72	13.43	18.66	29.85	39.55	47.01	49.25	53.73	59.70	63.43	35.32
Online	Gradient		7.46	17.91	**25.37**	31.34	35.82	40.30	46.27	49.25	52.99	53.73	34.08
	Bayesian	λ	3	5	7	10	13	16	19	22	25	30	-
		Avg. # DCPs	1.00	2.03	2.57	2.77	4.11	8.20	35.01	23.25	26.85	26.34	-
		Prec.	0.75	1.49	2.24	3.73	6.72	26.12	82.84	79.85	84.33	85.07	-
		Norm. Prec.	0.75	0.73	0.87	1.35	1.64	3.19	2.37	**3.43**	3.14	3.23	-

Table 8
Average execution time for change point detection methods.

	Algorithms	Avg. Exec. Time per Sample (sec)		
		CASME II	MMI	YoutubeECD
Offline	BinSeg	0.0156	0.0056	0.0090
	Window	0.0077	0.0031	0.0050
	DynP	0.1190	0.0119	0.0337
	PELT	0.0859	0.0092	0.0223
	BottomUp	0.0236	0.0083	0.0137
Online	Gradient	**0.0002**	**0.0001**	**0.0001**
	Bayesian	0.4420	0.1534	0.2677

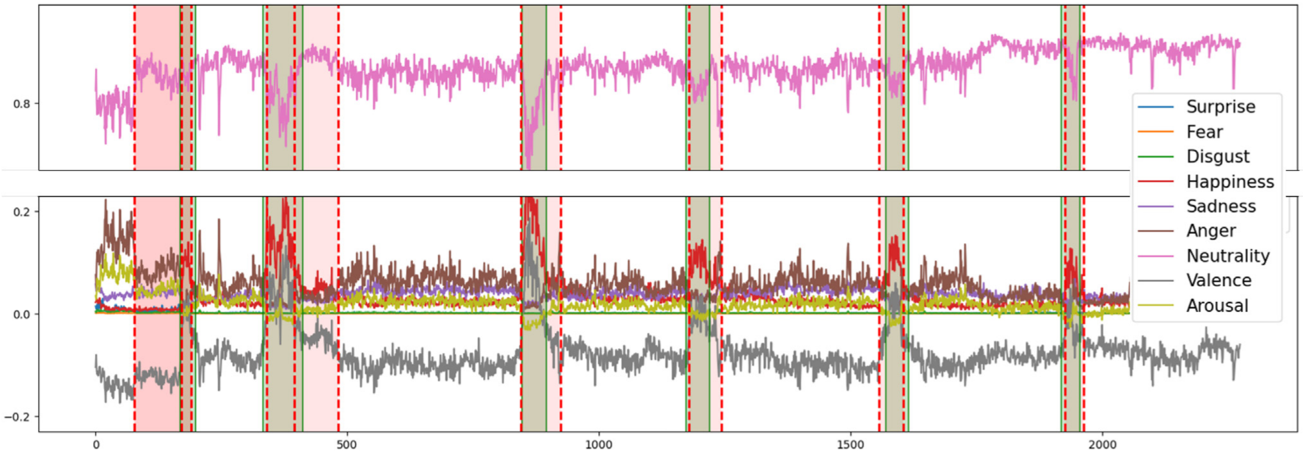


Fig. 8. The emotion change signal, ground-truth intervals (green shading), and predicted intervals (red shading) for the spotting task. The red dotted lines indicate the onset and offset points of the predicted intervals, whereas the green lines represent those of the ground-truth intervals. Note that the color darkens as the intervals overlap, so the first three red lines produced three predicted intervals. In this facial expression video, there are six true positives, three false positives, and no false negatives.

retrieved from a video sample of a person who had no facial expression for the most part but occasionally smiled for a short moment. By analyzing the emotion signal, this fact becomes apparent. Our method predicted nine spotted intervals although this video sample has six ground-truth intervals of facial expressions. Applying Eq. (24) to those results, there are 6 TPs, 3 FPs, and 0 FNs for this video sample.

Table 9 shows results of comparative experiments with the other spotting approaches on the CAS(ME)^2 dataset. The Yuhong et al. [83]' method produces the best f1-score among various spotting methods. However, this approach utilizes optical flow features to analyze facial movements on a strictly normalized face, indicating that it yields poor performance with facial pose variations. Even though our method was developed for the ECD task and trained without the spotting annotations, the modified version as described in Section 3.6, the spotting version of our Deep Emotion Change Detection (DECD-spot), shows the comparable f1-score with the second highest precision. This is interesting because there are 44 facial expression intervals with no offset like in Fig. 2 out of a total of 374 on CAS(ME)^2-cropped. That is, our DECD-spot does not find them because our method is designed to detect explicit changepoints for the ECD task. Furthermore, our DECD-spot does not use any training data from CAS(ME)^2-cropped. Nevertheless, our DECD-spot yields comparable results with the other methods. We also report Jun et al.'s approach [84] which is the first-ranked method at MEGC 2022 on CAS(ME)^3 dataset [85]. The DECD is placed fifth, despite the fact that it is not based on the assumption of the facial expression temporal model. Note that the DECD is the unmodified version whose outputs are composed of only a combination of the detected changepoints. This demonstrates that our framework is capable of identifying changes effectively. However, our approach cannot detect any micro-expressions since we need to raise the *pen* in Eq. (10) to maintain high detection performance. In other words, lowering the *pen* improves micro-expression detec-

Table 9
Comparisons of state-of-the-art algorithms for the temporal spotting task on the CAS(ME)^2 dataset. The CAS(ME)^2-cropped dataset is provided by MEGC 2021 for a fair evaluation. MaE and ME represent macro and micro facial expressions, respectively. DECD and DECD-spot depict the original version and the modified version of our DECD framework, respectively.

Algorithms	Dataset	Challenge	MaE	ME	Overall		
			F1-Score	F1-Score	Recall	Precision	F1-Score
Wang et al. [28]	CAS(ME)^2	-	-	-	-	-	0.0260
He et al. [78]	CAS(ME)^2	MEGC 2020	-	-	0.1196	0.0082	0.0376
Gan et al. [79]	CAS(ME)^2	MEGC 2020	-	-	0.1436	0.0098	0.0448
Pan et al. [27]	CAS(ME)^2	MEGC 2020	-	-	-	-	0.0595
Zhang et al. [80]	CAS(ME)^2	MEGC 2020	-	-	0.0547	0.2131	0.1403
Baseline [66]	CAS(ME)^2-cropped	MEGC 2021	0.0401	0.0118	-	-	0.0304
Pan et al. [81]	CAS(ME)^2-cropped	MEGC 2021	0.1250	0.0250	0.1597	0.0919	0.1168
Yang et al. [29]	CAS(ME)^2-cropped	MEGC 2021	0.2505	0.0153	0.1793	0.2310	0.2019
Yu et al. [82]	CAS(ME)^2-cropped	MEGC 2021	**0.380**	0.063	-	-	0.327
Yuhong et al. [83]	CAS(ME)^2-cropped	MEGC 2021	0.3782	**0.1965**	**0.5154**	0.2577	**0.3436**
Jun et al. [84]	CAS(ME)^3-cropped	MEGC 2022	-	-	0.4444	**0.2667**	0.3333
DECD (Ours)	CAS(ME)^2-cropped	-	0.2279	0.0000	0.2605	0.1816	0.2140
DECD-spot (Ours)	CAS(ME)^2-cropped	-	0.2675	0.0000	0.2353	0.2593	0.2467

tion but degrades overall performance. This point demonstrates our method's limitations in terms of micro-expressions, which we will address in future research.

5. Conclusion

In this study, we explicitly stated and formulated the facial expression-based ECD problem for the first time. The significance and necessity of the ECD problem were explained, and the justification for our problem was provided by demonstrating its distinction from the most similar problem in computer vision, the temporal spotting of facial expression. We designed a weakly supervised DECD framework that can be trained only with static facial expression images to detect the timing of emotion changes. Extensive experiments were carried out using our comprehensive DECD framework on various datasets with temporal labels, demonstrating the efficacy of our method for onset detection. Furthermore, our DECD framework was modified and tested for the spotting task, with comparable results, indicating that the ECD task can include the spotting task.

We conclude this paper with the hope that the facial expression-based ECD task will aid researchers in better understanding people's emotions, which are influenced by a variety of non-verbal means of communication by providing a crucial time for automatic emotion recognition. As our future work, we plan to expand the ECD task for multiple persons rather than a single person. Through the multi-person ECD task, it is feasible to examine whether an appropriate response occurred between participants at the proper time in social interactions. In particular, the multi-person ECD task can be utilized for automatic ASD screening because people with weak social communication capabilities, such as those with ASD, have trouble responding to others' emotions.

CRediT authorship contribution statement

ByungOk Han: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Cheol-Hwan Yoo:** Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Ho-Won Kim:** Validation, Formal analysis, Investigation. **Jang-Hee Yoo:** Conceptualization, Supervision, Funding acquisition. **Jinhyeok Jang:** Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Investigation, Supervision, Project administration.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: ByungOk Han reports financial support was provided by Korea Ministry of Science and ICT.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2019–0-00330, Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders based on Cognition of the Psychological Behavior and Response)

References

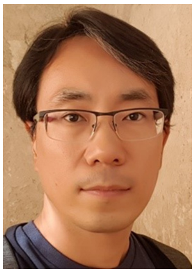
[1] A. Mollahosseini, B. Hasani, M.H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, IEEE Trans. Affect. Comput. 10 (1) (2017) 18–31.
[2] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, PloS one 9 (1) (2014).
[3] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: IEEE International Conference on Multimedia and Expo, 2005.
[4] M. Argyle, Non-verbal communication in human social interaction, Non-verbal communication 2 (1972).
[5] D. Phutela, The importance of non-verbal communication, IUP Journal of Soft Skills 9 (4) (2015) 43.
[6] M. Pantic, L. Rothkrantz, H. Koppelaar, Automation of non-verbal communication of facial expressions, in: European Conference on Media, Communication & Film (EuroMedia), 1998, pp. 86–93.
[7] A. Mehrabian, Nonverbal communication, Transaction Publishers, 1972.
[8] C. Frith, Role of facial expressions in social interactions, Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1535) (2009) 3453–3458.
[9] C.A. Corneanu, M.O. Simón, J.F. Cohn, S.E. Guerrero, Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications 38 (8) (2016) 1548–1568.
[10] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168–2177.
[11] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Trans. Affect. Comput. (2020).
[12] M.A. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, IEEE Trans. Affect. Comput. 2 (2) (2011) 92–105.

[13] J.A. Russell, A circumplex model of affect, J. Personality Social Psychology 39 (6) (1980) 1161.
[14] T. Jellema, A. Pecchinenda, L. Palumbo, E.G. Tan, Biases in the perception and affective valence of neutral facial expressions induced by the immediate perceptual history, Visual Cognition 19 (5) (2011) 616–634.
[15] L. Palumbo, T. Jellema, Beyond face value: does involuntary emotional anticipation shape the perception of dynamic facial expressions?, PloS one 8 (2) (2013)
[16] Y. Yamashita, T. Fujimura, K. Katahira, M. Honda, M. Okada, K. Okanoya, Context sensitivity in the detection of changes in facial emotion, Scientific Reports 6 (1) (2016) 1–8.
[17] S. Begeer, H.M. Koot, C. Rieffe, M.M. Terwogt, H. Stegge, Emotional competence in children with autism: Diagnostic criteria and empirical evidence, Developmental Review 28 (3) (2008) 342–369.
[18] E. Hill, S. Berthoz, U. Frith, Brief report: Cognitive processing of own emotions in individuals with autistic spectrum disorder and in their relatives, Journal of Autism and Developmental Disorders 34 (2) (2004) 229–235.
[19] O. FeldmanHall, T. Dalgleish, D. Mobbs, Alexithymia decreases altruism in real social decisions, Cortex 49 (3) (2013) 899–904.
[20] Z. Huang, J. Epps, E. Ambikairajah, An investigation of emotion change detection from speech, in: INTERSPEECH, 2015.
[21] Z. Huang, J. Epps, Detecting the instant of emotion change from speech using a martingale framework, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5195–5199.
[22] B. Han, H.W. Kim, J.-H. Yoo, Deep emotion change detection for human-robot interaction, in: International Conference on Intelligent Robots and Systems Workshops (IROSW), 2020.
[23] G. Sandbach, S. Zafeiriou, M. Pantic, D. Rueckert, Recognition of 3d facial expression dynamics, Image and Vision Computing 30 (10) (2012) 762–773.
[24] C. Izard, Basic emotions, relations among emotions, and emotion-cognition relations, Psychological Review 99 (3) (1992) 561–565, https://doi.org/10.1037/0033-295x.99.3.561.
[25] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, X. Fu, Cas (me) 2: a database for spontaneous macro-expression and micro-expression spotting and recognition, IEEE Trans. Affect. Comput. 9 (4) (2017) 424–436.
[26] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, IEEE Multimedia 19 (03) (2012) 34–41.
[27] H. Pan, L. Xie, Z. Wang, Local bilinear convolutional neural network for spotting macro-and micro-expression intervals in long video sequences, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020, pp. 749–753.
[28] S.-J. Wang, Y. He, J. Li, X. Fu, Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos 30 (2021) 3956–3969.
[29] B. Yang, J. Wu, Z. Zhou, M. Komiya, K. Kishimoto, J. Xu, K. Nonaka, T. Horiuchi, S. Komorita, G. Hattori, et al., Facial action unit-based deep learning framework for spotting macro-and micro-expressions in long video sequences, in: ACM Multimedia (MM), 2021, pp. 4794–4798.
[30] A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE Trans. Affect. Comput. 9 (1) (2016) 116–129.
[31] E.S. Page, Continuous inspection schemes, Biometrika 41 (1/2) (1954) 100–115.
[32] E. Page, A test for a change in a parameter occurring at an unknown point, Biometrika 42 (3/4) (1955) 523–527.
[33] J. Chen, A.K. Gupta, Parametric statistical change point analysis: with applications to genetics, medicine, and finance, Springer, 2012.
[34] K. Karagiannaki, A. Panousopoulou, P. Tsakalides, An online feature selection architecture for human activity recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2522–2526.
[35] D. Angelosante, G.B. Giannakis, Group lassoing change-points in piecewise-constant ar processes, EURASIP Journal on Advances in Signal Processing 2012 (1) (2012) 1–16.
[36] S. Chen, P. Gopalakrishnan, et al., Speaker, environment and channel change detection and clustering via the bayesian information criterion, in: DARPA Broadcast News Transcription and Understanding Workshop, Vol. 8, 1998, pp. 127–132.
[37] R. Killick, P. Fearnhead, I.A. Eckley, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association 107 (500) (2012) 1590–1598.
[38] A.G. Tartakovsky, B.L. Rozovskii, R.B. Blažek, H. Kim, Detection of intrusions in information systems by sequential change-point methods, Statistical Methodology 3 (3) (2006) 252–293.
[39] D. Kifer, S. Ben-David, J. Gehrke, Detecting change in data streams, in: VLDB, Vol. 4, 2004, pp. 180–191.
[40] R.P. Adams, D.J. MacKay, Bayesian online changepoint detection, arXiv preprint arXiv:0710.3742 (2007).
[41] Z. Wang, X. Lin, A. Mishra, R. Sriharsha, Online changepoint detection on a budget, in: IEEE International Conference on Data Mining Workshops (ICDMW), 2021, pp. 414–420.
[42] S. Niekum, S. Osentoski, C.G. Atkeson, A.G. Barto, Online bayesian changepoint detection for articulated motion models, in: IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 1468–1475.
[43] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: ACM on International Conference on Multimodal Interaction, 2015, pp. 503–510.
[44] B. Hasani, M.H. Mahoor, Facial expression recognition using enhanced deep 3d convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 30–40.
[45] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, L. Xun, Facial expression recognition with faster r-cnn, Procedia Computer Science 107 (2017) 135–140.
[46] P. Ekman, D. Keltner, Universal facial expressions of emotion, Nonverbal Communication: Where Nature Meets Culture (1997) 27–46.
[47] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, S. Lucey, Using synthetic data to improve facial expression analysis with 3d convolutional networks, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 1609–1618.
[48] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in: ACM International Conference on Multimodal Interaction (ICMI), 2016, pp. 445–450.
[49] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, C. Pal, Recurrent neural networks for emotion recognition in video, in: ACM International Conference on Multimodal Interaction (ICMI), 2015, pp. 467–474.
[50] Z. Yu, G. Liu, Q. Liu, J. Deng, Spatio-temporal convolutional features with nested lstm for facial expression recognition, Neurocomputing 317 (2018) 50–57.
[51] M. Nasir, P. Dutta, A. Nandi, Fuzzy triangulation signature for detection of change in human emotion from face video image sequence, Multimedia Tools and Applications 80 (21–23) (2021) 31993–32022.
[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[53] F. Ma, B. Sun, S. Li, Facial expression recognition with visual transformers and attentional selective fusion, IEEE Trans. Affect. Comput. (2021).
[54] Z. Zhao, Q. Liu, Former-dfer: Dynamic facial expression recognition transformer, in: ACM Multimedia (MM), 2021, pp. 1553–1561.
[55] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94–101.
[56] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2852–2861.
[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[58] W.H. Press, S.A. Teukolsky, Savitzky-golay smoothing filters, Computers in Physics 4 (6) (1990) 669–672.
[59] M.J. Orr et al., Introduction to radial basis function networks, Center for Cognitive Science, University of Edinburgh, 1996, Technical Report.
[60] C. Forbes, M. Evans, N. Hastings, B. Peacock, Statistical distributions, John Wiley & Sons, 2011.
[61] M.H. DeGroot, Optimal statistical decisions, Vol. 82, John Wiley & Sons, 2005.
[62] K.P. Murphy, Conjugate bayesian analysis of the gaussian distribution, Technical Report, University of British Columbia 1 ($2\sigma 2$) (2007) 16.
[63] P. Ekman, W.V. Friesen, Nonverbal leakage and clues to deception, Psychiatry 32 (1) (1969) 88–106.
[64] L. Zhou, X. Shao, Q. Mao, A survey of micro-expression recognition, Image and Vision Computing 105 (2021), https://doi.org/10.1016/j.imavis.2020.104043.
[65] P. Ekman, Telling lies: Clues to deceit in the marketplace, politics, and marriage, (revised edition),., WW Norton & Company, 2009.
[66] C.H. Yap, M.H. Yap, A.K. Davison, R. Cunningham, 3d-cnn for facial micro-and macro-expression spotting on long video sequences using temporal oriented reference frame, arXiv preprint arXiv:2105.06340 (2021).
[67] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[68] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision (ECCV), 2016, pp. 630–645.
[69] C. Truong, L. Oudre, N. Vayatis, Selective review of offline change point detection methods, Signal Processing 167 (2020).
[70] I. Radosavovic, R.P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10428–10436.
[71] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, X. Zhai, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.
[72] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: International Conference on Computer Vision (ICCV), 2021.
[73] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., Mlp-mixer: An all-mlp architecture for vision, Advances in Neural Information Processing Systems (NeurIPS) 34 (2021).
[74] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[75] C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, in: IEEE International Conference on Computer Vision (ICCV), 1998, pp. 839–846.

[76] J. Bai, Least absolute deviation estimation of a shift, Econometric Theory 11 (3) (1995) 403–436.
[77] M. Lavielle, Using penalized contrasts for the change-point problem, Signal Processing 85 (8) (2005) 1501–1510.
[78] Y. He, S.-J. Wang, J. Li, M.H. Yap, Spotting macro-and micro-expression intervals in long video sequences, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020, pp. 742–748.
[79] L. Jingting, S.-J. Wang, M.H. Yap, J. See, X. Hong, X. Li, Megc 2020-the third facial micro-expression grand challenge, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020, pp. 777–780.
[80] L.-W. Zhang, J. Li, S.-J. Wang, X.-H. Duan, W.-J. Yan, H.-Y. Xie, S.-C. Huang, Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020, pp. 734–741.
[81] H. Pan, L. Xie, Z. Wang, Spatio-temporal convolutional attention network for spotting macro-and micro-expression intervals, in: ACM Multimedia Workshops, 2021, pp. 25–30.
[82] W.-W. Yu, J. Jiang, Y.-J. Li, Lssnet: A two-stream convolutional neural network for spotting macro-and micro-expression in long videos, in: ACM Multimedia (MM), 2021, pp. 4745–4749.
[83] H. Yuhong, Research on micro-expression spotting method based on optical flow features, in: ACM Multimedia (MM), 2021, pp. 4803–4807.
[84] J. Yu, Z. Cai, Z. Liu, G. Xie, P. He, Facial expression spotting based on optical flow features, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7205–7209.
[85] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (3) (2022) 2782–2800.

ByungOk Han recieved his Ph.D. in Computer Science and his M.S. in Robotics from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016 and 2010 respectively. He received his B. S. in Computer Science and Engineering from Chung-Ang University, Seoul, South Korea, in 2008. Since October 2016, he has been with Electronics and Telecommunications Research Institute (ETRI), South Korea as a senior researcher. His research interests include computer vision, machine learning, pattern recognition, and human-robot interaction.

Cheol-Hwan Yoo received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2014 and 2020, respectively. Since 2020, he has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, as a senior researcher. His research interests include deep learning, image processing, computer vision, and human-robot interaction.



Ho-Won Kim received his Ph.D. and M.S. in Electrical and Electronic Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2004 and 1999 respectively. He received his B.S. in Electronics Engineering from Kyungpook National University, Daegu, Korea, in 1997. He worked as Senior Researcher in LG Electronics CTO Division, Korea in 2005. Since February 2006, He has been with Electronics and Telecommunications Research Institute (ETRI), Korea as a principal researcher. His research interests include computer vision, computer graphics, machine learning, AR/MR, digital human, and human-robot interaction

Jang-Hee Yoo received his BSc degree in physics and his MSc degree in computer science from Hankuk University of Foreign Studies, S. Korea, in 1988 and 1990, respectively. He received his PhD in electronics and computer science from the University of Southampton, UK, in 2004. Since November 1989, he has been with Electronics and Telecommunications Research Institute (ETRI), S. Korea as a principal researcher. He has also been a professor with the Department of Artificial Intelligence at the University of Science and Technology, S. Korea and was a visiting scientist at the University of Washington, Seattle, USA, from August 2014 to July 2015. His current research interests include computer vision, human motion analysis, biometric systems, HCI, and intelligent robot.

Jinhyeok Jang received the B.S. and M.S. degrees from the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea, in 2014 and 2016, respectively. Since 2017, he has been a Research Scientist with the Electronics and Telecommunications Research Institute, Daejeon, South Korea. His current research interests include image processing, blur estimation, human facial recognition, and human action recognition.