Received July 11, 2020, accepted July 26, 2020, date of publication August 11, 2020, date of current version August 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3015917

Complex Emotion Profiling: An Incremental Active Learning Based Approach With Sparse Annotations

SELVARAJAH THUSEETHAN, SUTHARSHAN RAJASEGARAR, (Member, IEEE), AND JOHN YEARWOOD, (Member, IEEE)
School of Information Technology, Deakin University, Geelong, VIC 3220, Australia
Corresponding author: Sutharshan Rajasegarar (srajas@deakin.edu.au)

ABSTRACT Generally, in-the-wild emotions are complex in nature. They often occur in combinations of multiple basic emotions, such as fear, happy, disgust, anger, sadness and surprise. Unlike the basic emotions, annotation of complex emotions, such as pain, is a time-consuming and expensive exercise. Moreover, there is an increasing demand for profiling such complex emotions as they are useful in many real-world application domains, such as medical, psychology, security and computer science. The traditional emotion recognition systems require a significant amount of annotated training samples to understand the complex emotions. This limits the direct applicability of those methods for complex emotion detection from images and videos. Therefore, it is important to learn the profile of the in-the-wild complex emotions accurately using limited annotated samples. In this paper, we propose a deep framework to incrementally and actively profile in-the-wild complex emotions, from sparse data. Our approach consists of three major components, namely a pre-processing unit, an optimization unit and an active learning unit. The pre-processing unit removes the variations present in the complex emotion images extracted from an uncontrolled environment. Our novel incremental active learning algorithm along with an optimization unit effectively predicts the complex emotions present in-the-wild. Evaluation using multiple complex emotions benchmark datasets reveals that our proposed approach performs close to the human perception capability in effectively profiling complex emotions. Further, our proposed approach shows a significant performance enhancement, in comparison with the state-of-the-art deep networks and other benchmark complex emotion profiling approaches.

INDEX TERMS Active learning, complex emotions, emotion recognition, incremental learning, sparse data.

I. INTRODUCTION

Humans convey their emotions as different nuanced expressions through their face. Although human regularly expresses continuous and complex emotions through face, preeminently, previous studies have mainly focused on detecting the Ekman's six basic emotions, namely happy, sad, surprise, fear, disgust and anger [1]–[3]. Accurately predicting the *complex emotions* (e.g., micro-emotions, pain and compound emotions) is essential to respond appropriately for a situation in many domains, such as medical, education and military. For instance, deceptive detection during a legal investigation is a good example for the use of extracted complex emotions, such as micro emotions. The detection of complex emotions assists an investigator to profile the subject's emotional state more precisely. Another significant application of complex emotion analysis is the detection of pain during a medical observation.

Du *et al.* [4] defines 21 categories of compound emotions that humans express in-the-wild. They have also demonstrated the correlations between those emotions and the action units (AUs), which are a group of muscles responsible for the facial expressions. For example, as illustrated in the top raw images of Figure 1, the compound emotion *happily surprised* is revealed by the presence of several muscle movements, namely AUs 1, 2, 5, 12, 25 and 26, as shown in the left image,

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

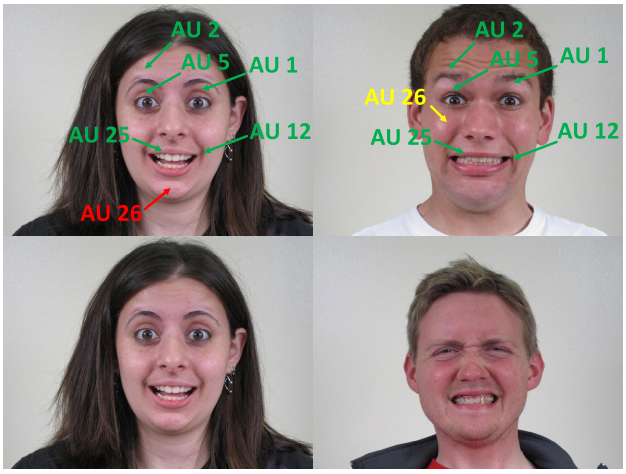

FIGURE 1. Examples of complex emotions. Top row: same complex emotions with different representations of AUs. Bottom row: Illustration of two complex emotions, happily surprised (bottom left) and happily disgusted (bottom right), that are formed from the combination of basic emotions happy, surprise and disgust.

and AUs 1, 2, 5, 6, 12 and 25, as shown in the right image. Further, the second row of images in Figure 1 shows two complex emotions, *happily surprised* (left) and *happily disgusted* (right), which are primarily formed as a combination of the basic emotions happy with surprise and disgust, respectively. On the other hand, definition of another complex emotion *pain* is provided in [5]. Micro-emotion is another significant complex emotion, which occurs for a fraction of a second, and hard to recognize in-the-wild using the naked eye.

The existing techniques for complex emotion profiling have *limitations in terms of detecting them* and often *require a large dataset* for training the emotion profiling model. Besides preliminary definitions, the complex emotions have not been analyzed deeply in the past due to two more significant reasons.

First, the complex emotions can only be extracted from a continuous observation in-the-wild, which is challenging due to the variations present in an uncontrolled environment, such as lighting and pose. Although in-the-wild facial analysis with highly uncontrolled environments has become the center of attention in the recent research studies, [6], [7], in-the-wild extraction of complex emotions has not been well addressed in the past. The applications, such as deceptive detection and pain estimation, require feedback in uncontrolled environments to enhance the quality of service. Hence an in-the-wild analysis of complex emotions demands an extensive examination.

Second, the unavailability or insufficient labeled datasets with annotations of complex emotions poses a significant limitation for the training of current deep networks. There are only a few benchmark datasets available with the annotations of complex emotions, such as micro-emotions, compound and pain emotions. Due to this limitation, the relatively new attempts on complex emotion recognition, such as pain intensity estimation, focused on using hand-crafted feature extraction techniques [8]. However, the minimal changes in facial muscles during the complex facial expression caused poor discriminative capability for the hand-crafted feature extractors. On the other hand, the rise of deep learning techniques (e.g., Convolutional Neural Network (CNN)) in the recent years enabled the computer vision systems to achieve highly efficient outcomes [9]. In addition, the deep CNN architectures have been widely utilized in recognizing the basic emotion through facial expressions. However, state-of-the-art deep learning techniques demand a large and balanced training dataset to perform optimally.

In order to address the research gap identified above, in this work, we propose a novel Active Hybrid Deep CNN framework with fusion mechanism, named as AHDCNN, to predict the complex emotions using facial expressions in-the-wild automatically. In AHDCNN, we introduce a cost-effective active learning (AL) based approach to improve the performance, and accelerate in-the-wild recognition of the complex emotion with a small amount of initial training data. Recent successes in AL-based approaches in computer vision provides motivations for complex emotion recognition with a small amount of annotated training data, which provides a less expensive way to train the model. AL is capable of providing a competitive classifier with a small number of initial training samples integrated with a progressive learning process in various image classification problems in-the-wild [10]. Further, the recently emerged AL approaches also demonstrated the reduced cost of labeling for training instances and improved performances [10]–[15]. However, integrating AL into deep architectures for image classification problems is limited due to the challenges, such as *unavailability of techniques to define the optimal size of initial training data for deep network architectures* and *inefficient active selection algorithms*.

Inspired by these two practical issues of integrating AL with deep network frameworks, we propose an enhanced AL technique that optimizes the initial training dataset. In particular, we utilize an image augmentation process. In addition, we propose an improved active selection algorithm that incorporates a wide range of samples ranging from informative to non-informative stage in the model updating process. We then propose an image pre-processing method to alleviate the variations present in uncontrolled environments in-the-wild. A variety of image pre-processing tasks have been proposed in the past. However, those conventional image pre-processing approaches have limitations for emotion profiling tasks due to *not being fine-tuned on more specific facial emotions* and *low robustness of the pre-processing tasks of image processing for unknown, in-the-wild, environments*. Motivated by these two image pre-processing related issues, we propose a comprehensive image pre-processing technique for in-the-wild facial emotion profiling task. This image pre-processing task is integrated as an internal component of our proposed framework.

A. KEY CONTRIBUTIONS

In summary, the key contributions in this paper are as follows:

- First, we develop an incremental active learning-based end-to-end deep CNN framework that performs accurate in-the-wild prediction of various complex emotions, such as micro-emotions, pain and compound emotions. In our deep framework, we introduce an improved cost-effective AL mechanism with a continuous and fully automated feedback mechanism. Moreover, our end-to-end framework is capable of estimating the optimized emotion dataset for initial training.
- Second, we propose a comprehensive image pre-processing mechanism, which is specifically designed for facial emotion images, to handle the inconsistency of an uncontrolled environment.
- Third, we show that the proposed framework yields state-of-the-art emotion prediction accuracies with small training sets in profiling the complex emotions in-the-wild. To validate this, we have compared the prediction accuracy with existing complex emotion recognition methods discussed in the literature and other five fine-tuned state-of-the-art deep networks.

The remainder of this paper is structured as follows. In *Section II*, the preliminaries of AL approach and complex emotions are reviewed. The methodology is introduced in *Section III*. Then, the *Section IV* describes the extensive experiments and evaluation. Lastly, *Section V* provides the conclusion and future directions of our work.

II. RELATED WORK

In this section, we describe the recently proposed related works on complex emotion profiling and active learning techniques.

A. COMPLEX EMOTION PROFILING

Apart from the basic emotions, humans express many complex emotions during continuous conversations. Although most of the existing works have focused on six basic human emotions, a list of complex emotions, such as pain, micro-emotions and compound emotions have been identified in the past due to its significance in many applications, such as medical interventions, human-computer interaction, sociable robots and social conversations.

1) COMPOUND EMOTIONS

Compound facial emotions are formed from a combination of a few existing basic emotions (e.g., happily surprised is a combination of basic emotions happy and surprise). Du *et al.* [4] defined 22 emotion categories, including the six basic emotions with neutral and 15 compound emotions. These compound emotions have not been analyzed in-depth using deep learning approaches due to the insufficient amount of labeled data to train the model. In particular, with 10-fold cross-validations, the authors of [4] have achieved classification accuracies of 73.61%, 70.03% and 76.91% when using shape, appearance and combined features respectively. Similarly, for the leave-one-out cross-validation, 72.09%, 67.48% and 75.09% classification accuracies were reported for shape, appearance and combined features. During the comparison, authors have reported that their shape and appearance-based model outperformed the multi-class SVM proposed in [16]. However, due to insufficient labeled data, authors have not compared the performance of their model with any of the existing deep networks.

2) MICRO EMOTIONS

Micro-emotion appears for a short duration with low intensity, which is also considered as one of the complex emotions since it is difficult to recognize in-the-wild. Numerous handcrafted feature extraction techniques have been proposed in the past to recognize the micro-emotions from videos. However, in recent years, a few prominent research works such as [17]–[19] have shown potential improvements in micro-emotion recognition using deep techniques.

In [17], authors have used a dual temporal scale CNN architecture to recognize the micro-emotions spontaneously. To avoid overfitting while training the deep model with sparse dataset, a dual architecture has been constructed based on two shallow CNN networks. Further, to acquire higher-level features, authors have used the optical flow frames instead of raw images. The experimental results show that the proposed architecture achieved 10% better accuracy than the state-of-the-art techniques. In [18], another significant study on micro-emotion recognition is presented, where the authors have utilized an enriched long-term RCNN. In this approach, the CNN modules are used to extract the features, and a long short-term memory (LSTM) is used to predict the micro-emotions. This approach also outperformed existing micro-emotion recognition techniques. However, the approaches proposed in [17] and [18] were not tested with sparse raw image training samples to recognize the micro-emotions.

Peng *et al.* [19] have proposed a transfer learning-based approach to recognize the micro-emotions considering a small training data. The ResNet10 [20] deep network that was pre-trained on Imagenet [21] dataset has been used to transfer learn on a small micro-emotion dataset. This approach achieved prediction accuracy rates of 70.59% and 75.68% on SAMM [22] and CASME II [23] datasets, respectively. Apart from the fact that it is working well with small datasets, a major limitation of this approach is its poor prediction accuracies compared to existing state-of-the-art micro-emotion recognition techniques.

3) PAIN

Highly social species, including humans, use face to express emotional states, such as pain during social and medical interaction. In the past, researchers have mainly focused on classifying the pain into binary classes, namely having pain or not. However, a vast number of recent research have focused on estimating the intensity of pain at a fine-grained level rather than a simple twofold classification. Facial action coding

system (FACS) provides a standard way of defining the pain intensity estimation, where FACS represents a movement of facial components based method, which is effective to represent emotions with rich expression states, such as pain. Numerous researchers have used FACS to estimate the pain intensities in the past. However, the Prkachin and Solomon Pain Intensity (PSPI) [5] metric has been widely used to estimate the pain intensities in a sixteen-level ordinal scale from a combination of six action units (AUs).

The majority of the existing pain estimation approaches are based on typical handcrafted feature-based techniques [8], [24], [25]. The lack of labeled data and standard rules cause the automatic feature extraction based pain intensity estimation challenging. Due to the limited deviations in painful facial expressions between subsequent PSPI scales, researchers tend to curtail the number of pain intensity classes to improve better detection performances. One notable work carried out by Hammal and Cohn [8] estimated the pain intensity into four levels using a handcrafted feature extraction method; defined as PSPI $= 0$ (none), PSPI $= 1$ (trace), PSPI $= 2$ (weak) and $PSPI \geq 3$ (strong). In this work, canonical appearance (C-APP) derived from the active appearance model (AAM) is traversed through Log-Normal filters in order to extract the features to classify the pain intensity classes. Additionally, four separate support vector machines (SVMs) have been trained using both 5-fold and leave-one-out cross-validation techniques. The classification rates achieved for the 5-fold and leave-one-out cross-validations are (97, 61), (96, 72), (96, 79), (98, 80) for the pain intensity levels none, trace, weak and strong respectively.

Roy *et al.* [24] designed another novel framework to estimate the pain intensity levels in four classes, as defined in [8]. A Gabor filtering was used for feature extraction, and Principal Component Analysis (PCA) was applied for feature compression. An SVM is then used to classify various pain intensity levels. The experiment was carried out under the frame level and image level settings in order to verify the robustness and accuracy of the framework under person dependent and person independent environments, respectively. Results show that this framework achieved 82.43% average classification accuracy over the four-level pain intensities. In [26], authors have categorized the pain intensities into six meaningful levels, namely none, mild, discomforting, distressing, intense and excruciating. Zhao *et al.* [25] studied estimation of the same six-level pain intensities, which are defined in [26]. The maximum estimation accuracy was achieved under a supervised setting among the experiments performed under fully supervised, semi-supervised and unsupervised settings. As observed, in [8], [24] and [25], the classification rate obtained for leave-one-out is significantly low, which is identified as a major limitation. It leads to a generalization issue for the proposed models, which is not effective across a range of different datasets.

Although the deep learning approaches have shown promising results during the recent years in various applications including computer vision, only a few works have been performed using automatic feature extractive deep learning techniques, especially in the automatic pain detection area [27]–[29]. In most cases, a limitation observed is the unavailability of annotated data for distinct pain intensity levels. By addressing this limitation, deep techniques still achieved comparable performances in this domain. In [27], Martinez *et al.* proposed a Recurrent Neural Network (RNN) based approach to estimate the pain intensity using the visual analog scale (VAS). A Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) [30] was used as the core of this model. Although it has achieved higher accuracies on lower intensity levels, the limited training data caused poor average accuracies on higher intensity levels. Wang *et al.* [28] and Zhou *et al.* [29] have recently attempted to use recurrent CNN to estimate the pain intensity automatically. In [28], authors fine-tuned a pre-trained network to transfer the knowledge for pain estimation. Conversely, in [29], a five-layer convolutional network has been used to train the system. In both cases, the proposed models were trained using the whole pain dataset and showed low classification accuracy.

B. ACTIVE LEARNING

Recent deep learning-based architectures heavily rely on accurately annotated training datasets to learn accurate models. In particular, as discussed before, complex emotions are challenging to be annotated, which leads to insufficient annotated training data. In order to mitigate the aforementioned problem, AL techniques have been proposed for use in a range of computer vision tasks. AL-based models are usually trained with sparse data, and actively improved using most informative samples. Various AL algorithms in conjunction with deep networks have been proposed in the past for vision tasks [10], [31]–[34].

Most of the existing works consider only the most informative or minority samples after an active user labeling is performed [32], [33], adopting common AL methods, such as *least confidence*, *margin sampling* and *entropy*. In [31], Li *et al.* proposed an adaptive AL framework that considers an uncertainty measure and a density measure to select the critical samples. This approach also failed to consider the majority or high confidence samples. In contrast, Wang *et al.* [10] presented a cost-effective active learning algorithm for deep image classification tasks that selects both majority and minority samples during the active selection process. This approach again depends on a costly and time-consuming human labeling process. Huang *et al.* [34] proposed a slightly different framework that considers two novel criteria, namely distinctiveness and uncertainty. Although the extensive experiments reported that this framework outperformed other active model adaptation techniques, a limitation is that it only considered binary classification. Further, their approach did not consider a diverse range of samples from unlabeled data. Thus the AL algorithm will fail to contemplate the majority of the samples during the selection process. However, previous studies have shown that considering

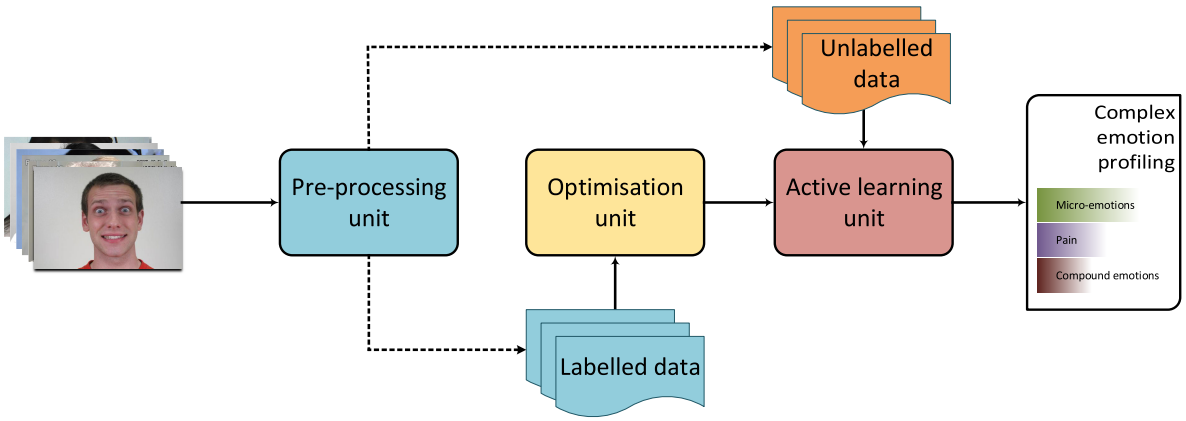

FIGURE 2. The proposed active learning based architecture to profile complex emotions.

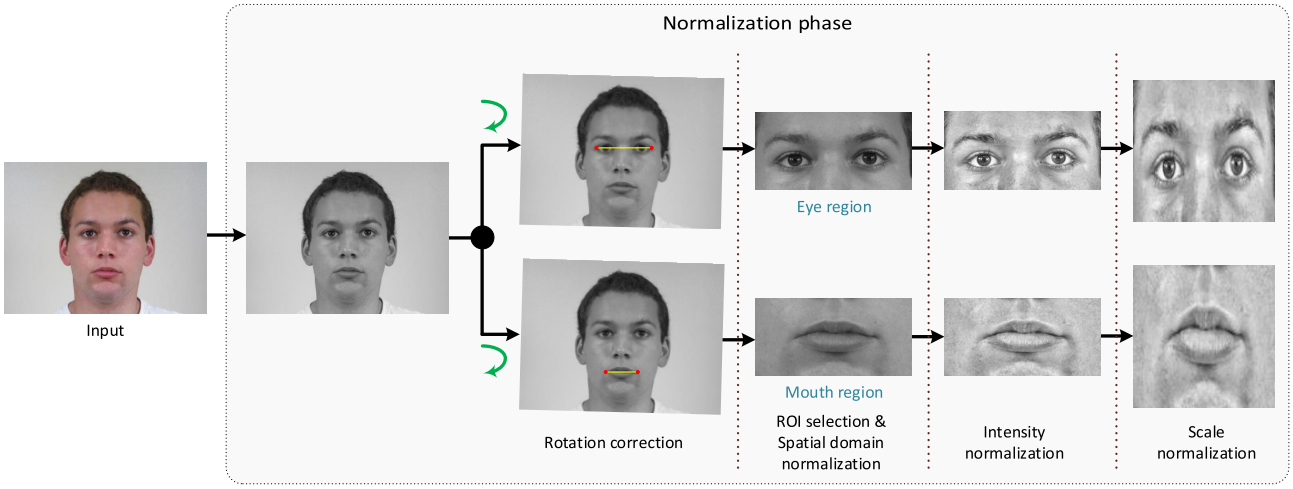

FIGURE 3. Pipeline of normalization steps included in the pre-processing unit in our framework.

samples with different confidence values improves the prediction accuracy.

III. METHODOLOGY

Based on the challenges observed in the literature survey above, in this section, we propose a novel incremental AL-based deep framework for complex emotion profiling in-the-wild. The proposed framework consists of three components, namely pre-processing unit, optimization unit and an active learning unit, as illustrated in Fig. 2.

A. PRE-PROCESSING UNIT

In this section, we develop a comprehensive pre-processing mechanism, which is crafted specifically for complex emotion profiling tasks. *Normalization* and *augmentation* are two phases of our pre-processing descriptor. This pre-processing technique extends the one we proposed in [35].

1) NORMALIZATION

The complete overview of the normalization phase in the pre-processing unit is illustrated in Fig. 3. In the normalization step, first, the input video frames are converted into greyscale images in order to reduce the cross-database discrepancy between the video frames.

Rotation Correction: We then make two copies of each greyscale image to perform rotation corrections. This process eliminates the rotation variation related complexity while extracting the features, thus providing a reliable way to extract emotion features from the face. During the rotation correction, as indicated in Fig. 3, we align the active appearance model (AAM) facial feature points 37 and 46 of eyes on the first image, and AAM facial feature points 49 and 55 of mouth on the second image horizontally. After that, first and second images are used to select the expression centric areas of the *eye* and the *mouth* regions, respectively, and are then spatially normalized.

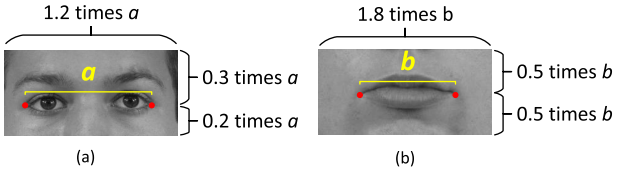

FIGURE 4. ROI selection and spatial domain normalization: (a) selected eye region (b) selected mouth region.

ROI Selection and Spatial Domain Normalization: Eliminating insignificant information (e.g., background information) in the input video frames will improve the detection or classification accuracy. The raw images of publicly available complex emotion datasets have a lot of background information in them. In the past, facial emotion recognition studies, such as [36], have eliminated the background information and certain portion of the face from the facial images to reduce the complexity. In our approach, not only the image is cropped to eliminate the background information and some portion of the face, but also the expression specific features are selected by focusing on the eye and mouth regions. The cropping process of eye and mouth regions are illustrated in Fig. 4.

For the eye region, we define a as the distance between AAM facial feature points 37 and 46. The width and heights are then set to 1.2 (0.1 times extended in each side) and 0.5 (0.3 times above and 1.1 below the eye corner AAM facial feature points) factors of a. Similarly, for the mouth region, b is defined as the distance between the lip corner AAM facial feature points 49 and 55. Then the width and heights are set to 1.8 (0.4 times extended in both left and right sides from the lip corner AAM facial feature points) and 1 (0.5 times above and 0.5 below the lip corner AAM facial feature points) factors of b. This is the average size of the active eye and mouth regions of all images used in the complex emotion datasets.

Intensity Normalization: The variations in image features, such as brightness and contrast often increase the complexity of classification tasks. Contrast limited adaptive equalization (CLAHE) [37] is one of the techniques that can be used to eliminate the variations in contrast and brightness of an image. CLAHE is a widely used variant of the adaptive histogram equalization algorithms, which can be applied on both colored and grayscale images. The slope of a transformation function, which is proportional to the cumulative distributive function (CDF) of neighborhood pixels, provides the contrast amplification of a pixel value. In CLAHE, before the computation of CDF, the contrast amplification is constrained by a pre-defined value called the clip limit of the histogram. The clip limit in CLAHE regulates the noise level that has to be smoothed, and the contrast that has to be enhanced. The primary advantage of CLAHE is that it redistributes the histogram part, which exceeds the clip limit between all histogram bins rather than just eliminating it. The clip limit

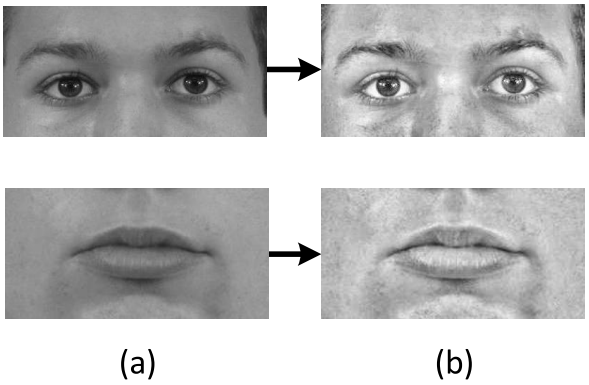

FIGURE 5. Illustration of the intensity normalization process applied on eye and mouth regions: (a) before and (b) after the intensity normalization process.

and the α value are set to 0.01 and 1, respectively, for the Rayleigh distribution used in this study. Fig. 5 illustrates an example of a sample image before and after the intensity normalization process.

Scale Normalization: As the last step of the normalization phase, we performed a scale normalization, where we down-sampled the size of the image to 128×128 pixels using linear interpolation. Scale normalization reduces the complexity of feature extractor by placing identical facial feature points of different images approximately at the same location.

2) DATA AUGMENTATION

Deep networks often show better performances with large training sets while performing classification tasks, such as profiling the complex emotions accurately. However, in this research, we have used a small portion of the benchmark complex emotion datasets for the training purposes. Therefore, we use synthetic data augmentation, which is often utilized to enhance the training set in the field of deep learning. Further, this technique has been widely used for many traditional deep network training purposes. Simard *et al.* [38] proposed a data augmentation method using elastic deformations (translation, rotation and skewing) on real images. Adopting this approach, we used a 2D Gaussian distribution to add random noise in the eye and mouth regions of the face to produce the synthetic frames separately. The Gaussian standard deviation is carefully engineered since both small and large variations can generate meaningless identical images and create a more complex learning environment for the classifier, respectively. Moreover, the augmented samples with large variations are carefully removed once again during the sample selection process. We synthesize all the rotation corrected images (both eye and mouth regions corrected images) and used to train the initial classifier.

B. SAMPLE SELECTION CRITERIA

In this section, we introduce the active sample selection criteria used in our framework. The main stages in sample

selection, namely confidence value calculation criteria, self pseudo labeling with high confidence samples and threshold fine-tuning are described below.

1) CALCULATION OF CONFIDENCE VALUE

In the past, many approaches have been proposed to calculate the *confidence* value using the probability of a predicted sample $P(y_i = j \mid I_i; \Omega)$ for a given deep CNN model Ω. Among them, three commonly used active learning techniques are least confidence [39], margin sampling [40] and entropy [41].

Culotta and McCallum [39] defined the least confidence criteria, which sorts the samples in an ascending order according to the classification probability predicted by the current model. Eq. 1 describes the definition of the least confidence criteria.

$$lc_i = \max_j(P_\theta(y_i = j \mid I_i; \Omega)) \tag{1}$$

where, $P(y_i = j \mid I_i; \Omega)$ indicates the classification probability of the sample I_i for the j^{th} class under the current model θ. The classifier is uncertain about a predicted sample when it records a lower confidence value.

Margin sampling [40] strategy, on the other hand, measures the confidence value according to the margin between the highest and the second-highest probable classes, as described in Eq. 2.

$$ms_i = P_\theta(y_i = j_{first} \mid I_i; \Omega) - P_\theta(y_i = j_{second} \mid I_i; \Omega) \tag{2}$$

where, $P_\theta(y_i = j_{first} \mid I_i; \Omega)$ and $P_\theta(y_i = j_{second} \mid I_i; \Omega)$ indicate the first and the second-highest classification probabilities of the sample I_i under the current model θ. The smaller margin indicates higher uncertainty of predicted sample by the current classifier. Thus, the samples are ranked in an ascending order.

Inspired by information theory, in entropy sampling [41] criteria, all the predicted class probabilities are utilized to measure the entropy, which is defined in Eq. 3. Higher entropy values for the predicted samples indicate the uncertainty of the current classifier. Hence, all the samples are arranged in descending order.

$$en_i = -\sum_{j=1}^{m} P_\theta(y_i = j \mid I_i; \Omega) \log P_\theta(y_i = j \mid I_i; \Omega) \tag{3}$$

where, en_i is defined as the summation of the probabilities of all possible classes (i.e., $j = 1 \ldots m$).

2) AUTOMATIC PSEUDO-LABELING

High confidence samples from the unlabeled dataset are selected to label the samples automatically, which are then included in the labeled set for the next training phase. We adopt the approach proposed in [42], where the authors have utilized least confidence, margin sampling and entropy criteria, in high-confidence sample selection for automatic pseudo-labeling.

Algorithm 1 Learning Algorithm for OSS

Input: Labeled samples L, which is a subset of the emotion dataset D. The optimization training set O. L^R and L^A are the reserved and augmented datasets respectively. S^R and S^A are the selected samples from the reserved and augmented datasets. The CNN parameters Ω that has to be optimized.
Output: The optimized CNN classifier's parameters Ω^O

procedure SAMPLE-SELECTION-OPTIMISATION(L, Ω)
  Initialise the CNN parameters Ω to Ω^I with the initial training set L^I
  while not reached the maximum training iterations **do**
    if $I_i \in \{L^R\}$ **then**
      Select a set of random samples S^R
    end if
    if $I_i \in \{L^A\}$ **then**
      Select the high confidence samples S^A using Eq. 4
    end if
    Add $S^R \cup S^A$ samples into the optimization dataset O
    Fine-tune the CNN parameters Ω^I to Ω^O using Eq. 6
    Update the selection threshold δ using Eq. 5
  end while
  return Ω^O
end procedure

$$j^* = \operatorname*{argmax}_j(P_\theta(y_i = j \mid I_i; \Omega))$$
$$y_i = \begin{cases} j^*, & en_i < \delta \text{ or } ms_i, lc_i > \delta. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In equation 4, j^* is the most probable label of the sample I_i with the current model. y_i describes the label with the highest prediction probability, where lc_i, ms_i and en_i are calculated using equations 1, 2 and 3 respectively.

The classification ability of the model incrementally grows during the active learning process. Thus, the selection threshold needs to be updated to improve the model's reliability with newly added labeled data. We update the selection threshold δ using the equation 5.

$$\delta = \begin{cases} \delta_0, & t = 0. \\ \delta - d_r * t, & \text{otherwise.} \end{cases} \tag{5}$$

where, the threshold δ is initially set to δ_0 and updated in each iteration using a learning rate decay d_r.

In the next subsection, we provide details about the optimization unit proposed in our framework.

C. OPTIMIZATION UNIT

After obtaining the pre-processed data, during the optimization phase, we train the initial optimized model, as illustrated in Fig. 6. To optimize the deep CNN model, we use the

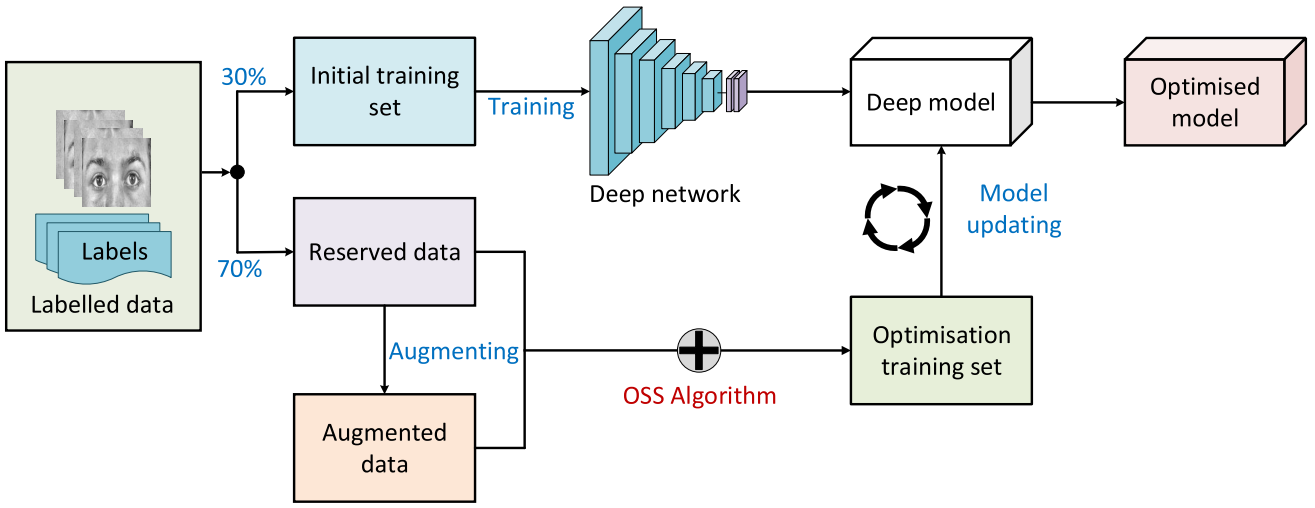

FIGURE 6. Illustration of the optimization unit. The optimization unit uses the optimization sample selection (OSS) algorithm to obtain an optimized model.

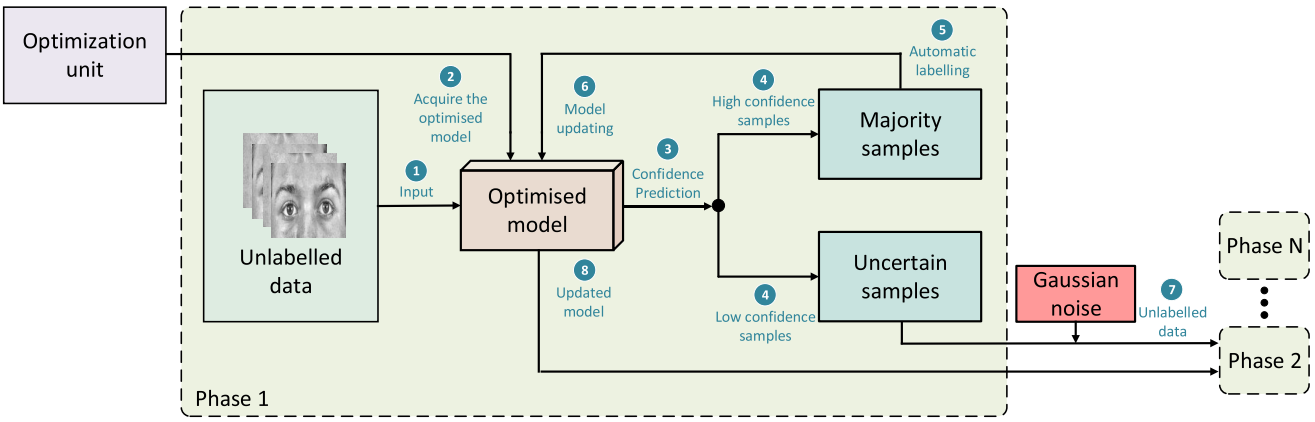

FIGURE 7. Active learning unit that uses the proposed active sample selection algorithm.

labeled dataset L, which is a subset of a given complex emotion dataset D (i.e., $L \subset D$). Initially, the deep CNN network is trained using 30% of the labeled data, which is $L^I(0.3 \times L)$, to initialize the parameters of the deep CNN parameters to Ω^I. The rest of the annotated dataset L^R $(0.7 \times L)$ is reserved for model optimization. The samples of the initial training dataset are randomly selected from the labeled data. After the initialization step, as indicated in Eq. 6, the model is updated incrementally using the optimization training set O, which is obtained from a combination of selected reserved L^R and augmented L^A samples.

$$\Omega^O = \omega(\Omega_i^I + \varepsilon_i O_i) \tag{6}$$

In Eq. 6, Ω^O is the optimized model, Ω_i^I and $\varepsilon_i O_i$ are the initial model and the optimizing weights in i^{th} iteration of the incremental process, where $i = 1 \ldots n$.

We propose a robust sample selection algorithm that can progressively select the samples from the optimization training set for the incremental model updating process. Algorithm 1 explains the steps involved in the optimization sample selection (OSS) algorithm in detail. The reserved data instances of the optimized training set are picked in the model updating process without any conditions. However, from the augmented portion of the optimization training set, only the majority of samples with high prediction confidence (i.e., clearly classified) have been selected for the incremental model updating process. This mechanism helps eliminate the augmented images that are highly deviated from the original images. Generally, augmented samples with high deviation increase the complexity of the deep classifiers.

In the OSS algorithm, the active user participation is not required to select samples from both reserved and augmented datasets. We only use previously annotated data to optimize the model. Hence, the selection of random samples from the reserved dataset is entirely based on the available annotations. Another advantage of OSS is the elimination of active

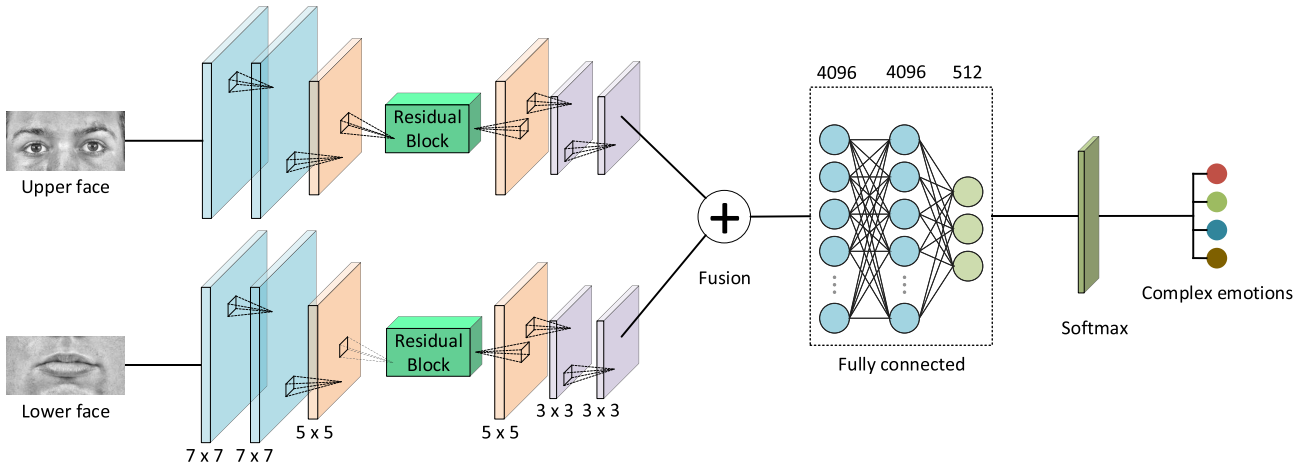

FIGURE 8. The proposed CNN architecture in our incremental active learning based deep framework.

Algorithm 2 ASS Algorithm

Input: Unlabeled samples U, which is a subset of the emotion dataset D. The CNN parameters Ω^O that need to be fine-tuned.

Output: The fine-tuned CNN classifier's parameters Ω^F

procedure ACTIVE-SAMPLE-SELECTION(U, Ω^O)
  while not reached the maximum training iterations **do**
    Select the high confidence samples U^H using Eq. 4
    Fine-tune the CNN parameters Ω^O to Ω^F using U^H
    Update the selection threshold δ using Eq. 5
    $U = G(U - U^H)$
  end while
  return Ω^F
end procedure

user participation through a high confidence sample selection technique to obtain the training samples from the augmented dataset. Thus, OSS completely eliminates the expense of active user participation in the optimization process.

D. ACTIVE LEARNING UNIT

The main purpose of proposing an active learning unit in our framework is to actively learn and enhance the classification capability of the obtained optimized model with minimum training data. Fig. 7 shows the proposed active learning unit, which uses a comprehensive active sample selection (ASS) algorithm. The proposed ASS algorithm, which is illustrated in Algorithm 2, utilizes the majority samples (i.e., clearly classified samples with high confidence values) in each iteration, like used in other conventional AL approaches. The intuition behind this algorithm is to select the samples with high confidence values to automatically annotate and add them into the training set. However, our approach additionally considers the minority samples (i.e., informative samples) in subsequent phases during an incremental model updating process. As illustrated in phase 1 of Figure 7, we use the optimized model Ω^O, which is derived from the optimization phase, to select the majority samples from the unlabeled data U. We then utilize the majority samples U^H to update the model. We then add Gaussian (G) noise to the minority samples $U - U^H$, and reserve the resultant samples $G(U - U^H)$ to present as an input for the next phase along with the updated model. Subsequently, we update the selection threshold using Eq. 5. We repeat the aforementioned incremental based model updating process until there is no further significant improvement in learner performance is observed, i.e., the training loss of the classifier is converged.

Next, we explain the deep CNN architecture used in our approach.

E. DEEP NETWORK

The novel CNN architecture integrated into our framework is illustrated in Figure 8. Our proposed deep network architecture consists of two parallel CNN stacks, each with six convolution layers, which is shallower than the majority of the existing state-of-the-art deep networks. Since the training process starts with small complex emotion datasets, in both optimization and active stages, using very deep networks are vulnerable for overfitting. Hence, as indicated in the figure, we have chosen a network with fewer convolution layers with appropriately placed residual blocks, where, each adding six extra convolution layers to our network. Residual blocks ultimately increase the number of layers in the network while providing flexibility to skip the training of a few convolution layers, and hence minimizing the complexity of the deep network. Generally, the skipped connections in the residual block eliminate the degradation problem during the training phase. In addition, after each convolution layer in our primary network, multiple rectified linear units (ReLU),

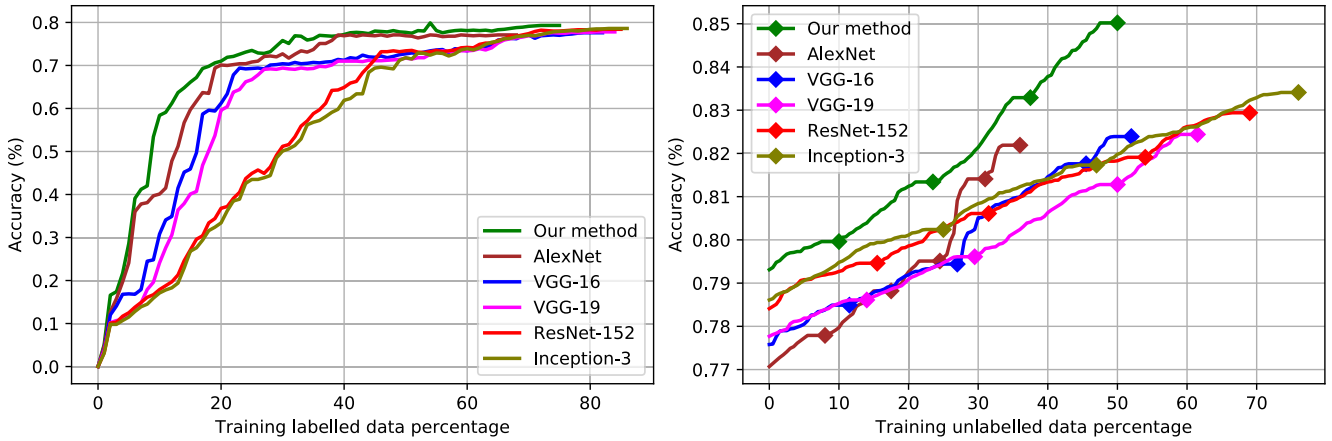

FIGURE 9. Comparison between state-of-the-art deep networks and our approach for optimization and active learning stages on Compound dataset (use the color image online for better viewing).

dropout, normalization and pooling are attached to improve the stability of the deep networks.

In summary, we configure both stacks of our parallel deep CNN identically. Each stack of the network accepts images of size 128 × 128 with 3-channels, where the upper stack accepts the upper face and the lower stack accepts the lower face, as illustrated in Figure 8. The first two convolution layers are implemented with a kernel of size 7 × 7, a stride of size 2 and a padding of size 1. The kernel size and the stride size are set to 5 × 5 and 1 for the third and the fourth convolution layers. For the last two layers, the kernel size is further reduced to 3 × 3 with the stride size of 1. For the last four layers, the padding is set to 0. One ReLU layer is always attached immediately after each convolution layer of our primary network. There are two dropout layers placed after the third and fifth convolution layers. Additionally, other than the first convolution layer, a pooling layer is placed after every other convolution layer in our architecture. After fusing the feature maps, we stack 3 fully connected layers with sizes 4096, 4096 and 512, respectively. The first two fully connected layers are followed by two dropout layers in the network.

As indicated earlier, a residual block is placed between the third and fourth convolution layers of the main network, which comprised of 3 skip connections. The first and last of the six convolutional networks implemented in the residual block are with the kernel size of 5 × 5 and stride size of 2. The kernel size and the stride size of the rest of the convolutional networks are set to 3 × 3 and 1. The padding size is 0 for all the convolution layers in the residual block.

Finally, a softmax layer is utilized to perform the complex emotion classification.

IV. EXPERIMENTS AND RESULTS

In this section, we present the results and analysis of the extended experiment carried out on publicly available complex emotion benchmark datasets, to demonstrate the cost-effectiveness of our proposed active incremental learning approach. We report the results separately for three different types of complex emotions, namely compound, micro-expressions and pain, which are discussed before.

TABLE 1. A summary of the complex emotion datasets used to evaluate the proposed approach.

Dataset	Emotion type	Samples	Classes
Compound [4]	Compound	230	21
CASME [43]	Micro	195	8
CASME II [23]	Micro	247	5
CAS(ME)^2 [23]	Micro/Macro	303	4
SAMM [22]	Micro	159	7
UNBC [44]	Pain	200	1

A. EXPERIMENTAL SETTING

1) COMPLEX EMOTION DATASETS

Here, we evaluate the proposed deep active learning-based approach and report the results on various complex emotion datasets, such as the *compound emotion dataset* [4], *micro expression datasets* CASME [43], CASME II [23], CAS(ME)^2 [62] and SAMM [22], and the *pain dataset* UNBC-McMaster Shoulder Pain Expression Archive (UNBC) [44]. Table 1 illustrates the summary of complex emotion datasets used in our experiments. The compound emotion dataset was collected from 230 subjects, which provides annotation of 21 emotion categories, which includes 6 basic and 15 compound emotions. The CASME [43] is claimed to be the first spontaneous micro emotion dataset that provides annotations for 8 micro emotions, such as amusement, sadness, disgust, surprise, contempt, fear, repression and tense. The authors further extended the dataset to CASME II [23], which provides much more sophisticated annotations for 5 micro-expression (i.e., happiness, disgust, surprise, repression and others).

The CAS(ME)^2 [62] is another spontaneous dataset that offers 303 expression samples, including 53 micro-expression sequences of four classes, such as positive, negative, surprise and other. Meanwhile, Davison *et al.* recently dispensed another spontaneous micro-expression dataset, namely SAMM [22], that contains annotated samples for 7 emotion classes, including 6 basic emotions. Lastly, the UNBC pain dataset is dedicated for the emotion pain and its intensity levels. The UNBC pain dataset consists of 200 video sequences with frame-level pain intensity annotations.

In order to perform a fair evaluation with existing complex emotion recognition methods, we use 10-fold and leave-one-subject-out (LOSO) cross-validation techniques to report our results. In both cross-validation techniques, labeled, unlabeled and test sets are manually sliced and consistently swapped across the whole dataset samples. For the 10-fold cross-validation, in each iteration, 10% samples of the whole dataset are reserved as the test set to report the performance of our model. Additionally, in each complex emotion dataset, we reserve 30% of the annotated samples as the labeled portion for the model optimization purpose, and the rest as the unlabeled portion for the active learning process. In contrast, for the LOSO protocol, we reserve one subject for testing purposes in each iteration and present the average results. We followed the similar protocol that we used for the 10-fold cross-validation to generate the labeled and unlabeled sets for optimization and active learning purposes.

2) IMPLEMENTATION

In the training phase, stochastic gradient descent with momentum (SGDM) method is used as the optimizer. Other parameters, such as learning rate, momentum, weight decay and Gaussian standard deviation are set to 10^{-6}, 0.9, 5×10^{-5} and 10^{-2} respectively. The same CNN parameter values are used without any changes in the experiments carried out on all complex emotions datasets. For the active learning environment, we set the initial threshold δ_0 and decay rate d_r as shown below in pairs, for least-confidence, margin-sampling and entropy-based methods respectively: [(8×10^{-1}, 0.2×10^{-6}), (8×10^{-1}, 0.2×10^{-6}) and (0.2×10^{-6}, -0.1×10^{-6})]. These parameters are updated throughout the training process.

3) METRICS

The metrics used in complex emotion analysis are *accuracy*, *mean squared error (MSE)* and *Pearson's product-moment correlation coefficient (PCC)*. The accuracy is used to present majority of the results in our experiments, which is defined in Eq. 7.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

where, TP, FP, TN and FN are *true positive*, *false positive*, *true negative* and *false negative*, respectively.

Additionally, for the pain intensity estimation experiment, MSE and PCC are gradually used to report the performance of our proposed approach, which are defined in Eq. 8.

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \\ PCC &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}_i})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}_i})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \end{aligned} \tag{8}$$

where, n is the number of samples in the test set. y_i and $\bar{y}$ are the ground-truth of the i^{th} frame and the mean of $\{y_1, \ldots, y_n\}$. $\hat{y}_i$ and $\bar{\hat{y}}$ are the predicted pain intensity level of the i^{th} frame and mean of $\{\hat{y}_1, \ldots, \hat{y}_n\}$, respectively. A higher value for PCC is better while a lower value for MSE is better.

B. EVALUATION FOR COMPOUND EMOTIONS

First, in this experiment, we evaluate the performance of our proposed framework on classifying the neutral face and 21 compound emotions defined in the compound emotion dataset [4]. Figure 9 illustrates the improvement of average classification accuracies of our approach over the five other state-of-the-art deep networks, considering the percentage of training data utilized in both optimization (left image) and active learning (right image) units. It can be seen that our approach performs favorably in both optimization and active learning steps against the compared deep networks. Our proposed framework has utilized 78% of the labeled training data for the model optimization to reach a stable average accuracy of 73.9%. It shows that the presented model is feasible, and can be optimized with a small labeled dataset. Other deep networks except for AlexNet [45], compared in this experiment, consumed more training labeled data for the model optimization. The proposed model also achieved a better accuracy in the optimization stage compared to the other state-of-the-art deep networks.

After the optimization, in the incremental active learning phase, the average classification accuracy of our approach has improved significantly and reached the peak average accuracy of 85.02% only with 50.5% of the unlabeled training data. It is clear that the incremental active learning phase has significantly improved the average classification accuracy of compound emotions. In addition, our model has recorded consistent accuracies ($\geq$ 72%) for each emotion as summarized in Table 2. As can be seen in the table, we compared the accuracies for each compound emotion with Du *et al.* [4] and five other state-of-the-art deep networks. For some emotional states, such as neutral, sadly fearful, fearfully angry, angrily surprised, angrily disgusted and hate, the best accuracy rates were not recorded by our model due to the fact that these emotion classes contain a considerable amount of intra-class variations that can easily be confused with other classes. However, the overall comparison results show that our model showed better classification ability for most of the compound emotions.

TABLE 2. Comparison of the average classification accuracy rates obtained for 22 emotion categories that includes neutral, six basic and 15 compound emotions using 10-fold cross-validation.

Methods	Neutral	Happy	Sad	Fearful	Angry	Surprised	Disgusted	Happily surprised	Happily disgusted	Sadly fearful	Sadly angry	Sadly surprised	Sadly disgusted	Fearfully angry	Fearfully surprised	Fearfully disgusted	Angrily surprised	Angrily disgusted	Disgustedly surprised	Appalled	Hate	Awed
Du et al. [4]	**.96**	.97	.87	.87	.77	.92	.81	.93	.91	.69	.85	.77	.77	.61	.54	.58	.71	.70	.83	.63	.65	.58
AlexNet [45]	.94	.91	.89	.84	.81	.89	.79	.95	.93	.81	.89	.81	.82	.74	.68	.71	.74	.82	.88	.74	.76	.69
VGG-16 [46]	.95	.92	.89	.86	.78	.91	.78	.95	.93	**.88**	.87	.83	.85	**.76**	.71	.69	.75	.74	.86	.73	.78	.71
VGG-19 [46]	.94	.94	.88	.87	.78	.92	.82	.94	.94	.78	.86	.83	.83	.74	.68	.66	**.76**	**.84**	.89	.75	.77	.72
ResNet-152 [20]	.93	.96	.89	.86	.76	**.93**	.83	.93	.94	.81	.85	.85	.84	.73	.69	.69	.75	**.84**	.87	.74	.80	.78
Inception-3 [47]	.92	.97	.90	.85	.78	.92	.85	.94	.96	.83	.84	.86	.83	.71	.71	.71	.74	**.84**	.88	.73	**.82**	.76
Our model	.90	**.98**	**.91**	**.89**	**.82**	**.93**	**.87**	**.96**	**.97**	.83	**.91**	**.87**	**.86**	.72	**.76**	**.78**	.73	.81	**.90**	**.76**	.74	**.80**

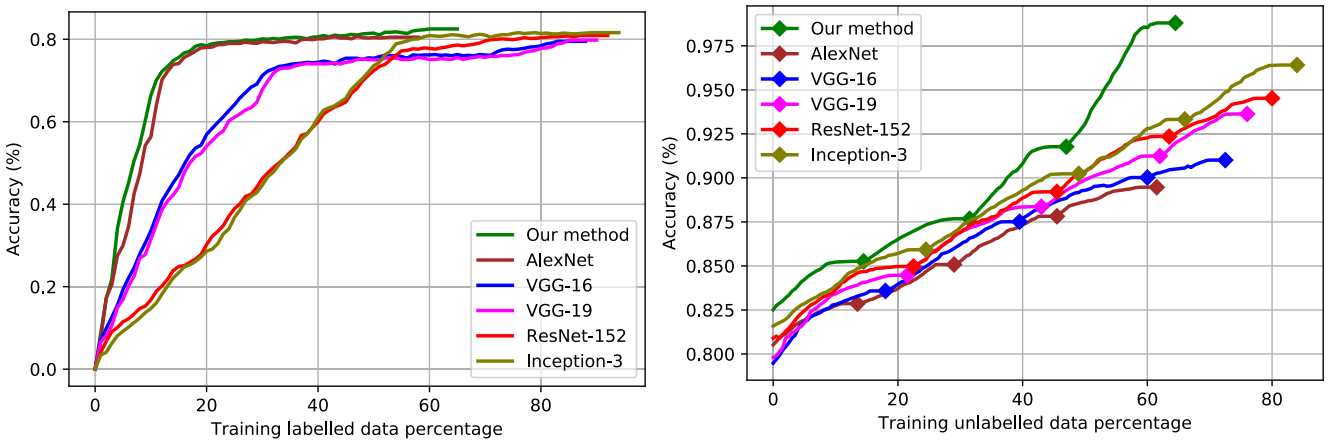

FIGURE 10. Comparison between state-of-the-art deep networks and our approach for optimization and active learning stages on UNBC dataset (use the color image online for better viewing).

TABLE 3. Comparison of the overall accuracy and F1-score achieved on Compound dataset [4] using both 10-fold and LOSO cross-validations.

	10-fold		LOSO	
Method	**Acc**	**F1**	**Acc**	**F1**
Du et al. [4]	.77	–	–	–
AlexNet [45]	.82	.69	.65	.60
VGG-16 [46]	.82	.71	.67	.63
VGG-19 [46]	.82	.72	.69	.64
ResNet-152 [20]	.83	.74	.71	.69
Inception-3 [47]	.83	.75	.75	.69
Our method	**.85**	**.81**	**.79**	**.76**

Further, the comparison of the overall results achieved on the Compound emotion dataset is shown in Table 3. It can be seen that our proposed model achieved better overall results on the Compound emotion dataset. Notably, our model shows a way better F1-score compared to other existing models, which shows that our model is effective with imbalanced datasets as well.

C. EVALUATION FOR PAIN INTENSITY ESTIMATION

Second, as described earlier, we evaluate the presented framework on UNBC pain dataset [44] for pain intensity estimation. In this experiment, we perform 16-level pain intensity estimation using the presented model, where the pain intensity levels are as defined in the PSPI metric. Figure 10 presents the comparison of average accuracy change against the percentage of the labeled data during optimization (left) and incremental active learning (right) stages by our model and other deep networks on the pain dataset. The results demonstrate that our proposed framework achieved 82.5% and 98.8% accuracies after optimization and incremental active phases, respectively, which is better than the state-of-the-art deep networks compared here. In addition, our approach used 66% and 65% of the labeled samples in the respective stages, which is much lower compared to the other deep networks, except for AlexNet [45].

TABLE 4. Comparison of pain intensity estimation results between the proposed approach and the other state-of-the-art methods along with the deep networks in the literature using UNBC pain [44] dataset.

	10-fold			LOSO		
Methods	**Acc**	**MSE**	**PCC**	**Acc**	**MSE**	**PCC**
Lucey et al. [44]	.84	–	–	–	–	–
Kaltwang et al. [48]	–	–	–	–	1.39	.59
Rathee and Ganotra [50]	.96	–	–	–	–	–
Hong et al. [49]	–	1.42	.55	–	–	–
Zhou et al. [29]	–	–	–	–	1.54	.65
Thuseethan et al. [52]	–	1.29	.73	–	–	–
AlexNet [45]	.90	1.56	.64	.68	1.67	.57
VGG-16 [46]	.91	1.58	.66	.71	1.61	.60
VGG-19 [46]	.94	1.46	.71	.77	1.54	.62
ResNet-152 [20]	.95	1.34	.70	.77	1.52	.62
Inception-3 [47]	.96	1.29	.72	.80	1.37	.66
Our method	**.99**	**1.21**	**.79**	**.87**	**1.26**	**.72**

Further, we compare our proposed framework with existing pain intensity estimation methods and the state-of-the-art deep networks in Table 4. The comparison shows that our approach outperforms the existing pain intensity estimation benchmark methods and the state-of-the-art deep networks by a comprehensive margin. Our method achieved the highest overall accuracy of 98.8%, MSE of 1.21 and PCC of 0.79 with 10-fold cross-validation. The low MSE reported for our method demonstrates that the majority of the misclassified samples are confused with nearby pain intensity classes. Some of the very recent pain intensity estimation methods, such as [51], are not compared with our approach as they use minimized pain intensity classes.

D. EVALUATION FOR MICRO EXPRESSIONS

Third, to demonstrate the model feasibility and effectiveness for the recognition of subtle micro-expressions, we further evaluated our presented framework on four benchmark micro expression datasets, namely CASME [43], CASME II [23], CAS(ME)^2 [62] and SAMM [22]. From the observation, the accuracy changes for our approach and other existing deep networks show a similar behavior as that are achieved in compound emotion [4] and UNBC pain [44] datasets. Compared to the state-of-the-art deep networks, our approach obtained the best accuracies in both optimization and incremental active learning stages on all three micro expression datasets.

Table 5 and 6 present the comparison of recent existing benchmark micro-expression recognition methods with our proposed approach. It can be observed that our approach outperformed all the state-of-the-art deep networks with 10-fold cross-validation. For the LOSO cross-validation, our approach outperformed all the benchmark micro-expression recognition approaches on CASME [43], CASME II [23] and CAS(ME)^2 [62] comprehensively. In addition, our approach has shown a competitive performance for the micro-expression recognition on SAMM [22] dataset. Yet, from the comparison, Thuseethan *et al.* [35] achieved a better performance on SAMM [22] dataset. However, in particular, our approach has utilized less amount of labeled samples for training compared to [35]. In comparison with our proposed approach, we can also see that the classification accuracy recorded for [35] is much lower on other two micro expression datasets.

In order to evaluate the generalization ability of our proposed framework for micro-expression recognition, a cross-database evaluation has been carried out on selected micro expression categories, which are commonly available (e.g., disgust) in all three datasets. To perform this, we trained our framework on one dataset and tested on other two. The corresponding results are presented in Table 7.

TABLE 5. Benchmarking against existing micro-expression recognition approaches and the state-of-the-art deep networks on CASME [43], CASMEII [43] and SAMM [43] datasets with 10-fold cross-validation.

	CASME		CASMEII		CAS(ME)^2		SAMM	
Methods	Acc	F1	Acc	F1	Acc	F1	Acc	F1
AlexNet [45]	.77	.72	.79	.73	.71	.65	.85	.76
VGG-16 [46]	.72	.69	.74	.66	.75	.72	.82	.74
VGG-19 [46]	.76	.72	.77	.71	.79	.76	.84	.73
ResNet-152 [20]	.81	.74	.80	.75	.81	.80	.88	.80
Inception-3 [47]	.86	.81	.82	.76	.84	.82	.87	.82
Our method	**.91**	**.87**	**.90**	**.87**	**.93**	**.92**	**.92**	**.87**

TABLE 6. Benchmarking against existing micro-expression recognition approaches and state-of-the-art deep networks on CASME [43], CASMEII [43] and SAMM [43] datasets with LOSO cross-validation.

	CASME		CASMEII		CAS(ME)^2		SAMM	
Methods	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Huang et al. [53]	–	–	.60	–	–	–	–	–
Huang et al. [54]	.57	–	.58	–	–	–	–	–
Zheng et al. [55]	.69	–	.63	–	–	–	–	–
Wang et al. [56]	–	–	.75	–	–	–	–	–
Zhang et al. [57]	–	–	.63	–	–	–	–	–
Zheng [58]	.71	–	.65	–	–	–	–	–
Thuseethan et al. [35]	.77	–	.82	–	–	–	**.91**	**–**
Verma et al. [59]	**.81**	–	.77	–	.76	–	–	–
Li et al. [60]	.54	–	.59	–	–	–	–	–
Song et al. [61]	.74	.73	.81	.81	.72	.71	.72	.69
Qu et al. [62]	–	–	–	–	.76	–	–	–
Reddy et al. [63]	–	–	–	–	.82	–	–	–
AlexNet [45]	.67	.61	.76	.68	.65	.61	.84	.70
VGG-16 [46]	.61	.54	.67	.60	.69	.61	.81	.66
VGG-19 [46]	.69	.63	.74	.65	.72	.65	.85	.77
ResNet-152 [20]	.72	.70	.76	.72	.76	.73	.88	.79
Inception-3 [47]	.79	.72	.78	.73	.80	.77	.87	.78
Our method	**.81**	**.79**	**.86**	**.82**	**.90**	**.86**	.89	**.83**

TABLE 7. Cross database evaluation.

Train dataset	CASME	CASMEII	CAS(ME)^2	SAMM
CASME	–	.81	.78	.75
CASMEII	.81	–	.76	.73
CAS(ME)^2	.77	.78	–	.70
SAMM	.68	.71	.69	–

The cross-database evaluation reveals that CASME [43] and CASMEII [23] are best generalized on each other, and demonstrated a less generalization on the SAMM [22] dataset. This is due to the fact that CASMEII [23] dataset is an extension of CASME [43], and both contain a part of the same samples. Moreover, this may follow an additional rationale that both CASME [43] and CASMEII [23] datasets were collected under the same environment, unlike SAMM [22] dataset, which was constructed under a completely different environment. The CAS(ME)^2 [62] dataset also achieved better accuracies on CASME [43] and CASMEII [23] datasets in comparison to SAMM [22] dataset. However, in summary, the classification accuracies obtained for the cross-database evaluation are satisfactory, and affirms that our model is readily generalizable.

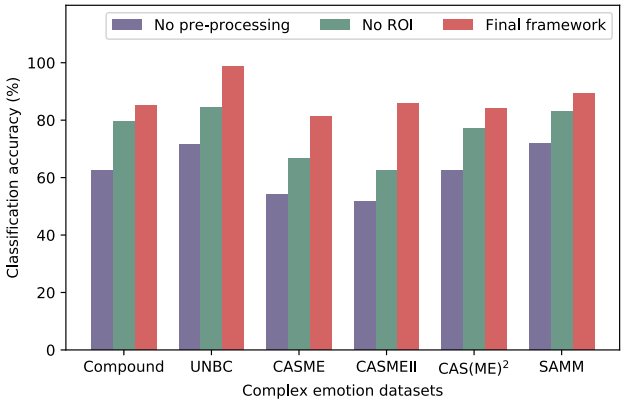

FIGURE 11. Comparison of the classification accuracies obtained in the ablative study.

E. ABLATIVE STUDY

Our proposed framework combines an image pre-processing unit, as described in Section 3.1. To justify that the integrated pre-processing technique improves the performance of the presented framework, we have carried out an ablative study. To perform this, we compare the classification accuracies after eliminating all or a few significant stages of our pre-processing phase (a) no pre-processing and (b) no ROI selection and spatial domain normalization with our (c) final framework, which includes all the pre-processing stages. The obtained accuracies of these variants on all complex emotion datasets are shown in Figure 11. The results clearly indicate that the integrated pre-processing unit considerably enhances the performance of our proposed complex emotion recognition framework. In particular, the classification accuracy has improved by 34.35% (i.e., from 51.67% to 85.97%) for the framework without the processing unit of the final framework on CASMEII [23] dataset. This substantial performance improvement confirms the significance of our pre-processing unit in our framework.

F. COMPUTATIONAL COMPLEXITY ANALYSIS

We further present a comprehensive comparison of the computation complexity between our approach and other state-of-the-art deep networks. The computing environment used to obtain the computation complexity results is Intel(R) Xeon(R) 2.20 GHz processor accelerated using NVIDIA GPU with GeForce GTX 1080 Titan. Figure 12 presents the training and the testing time consumed on each for the complex emotion datasets. To simplify the comparison, we first set the computational cost of our model to 1. Then, we represent the computational costs of other state-of-the-art deep networks as the number of times opposed to the computational cost of our method. As can be seen, our method is computationally efficient compared to existing state-of-the-art deep networks in recognizing complex emotions. In the training phase, our approach is 33.6% and 185.4% time efficient compared to AlexNet and Inception-3, respectively. In particular, compared to existing state-of-the-art deep networks, our approach is more effective in the testing process, as it shows 47.4% and 216.8% better time complexity against AlexNet and Inception-3.

G. DISCUSSION

In general, our extensive experiments on all three complex emotion scenarios show that our presented framework is promising compared to existing benchmark complex emotion methods and state-of-the-art deep networks. As a common pattern, it can be seen that the AlexNet [45] progressed better in the optimization stage, and achieved a competitive accuracy to our approach with a small amount of labeled samples. However, AlexNet [45] failed to converge in the incremental active phase to outperform our method. In contrast to AlexNet [45], more deeper networks such as ResNet-152 [20] and Inception-3 [47] showed slow progression in the optimization stage and converged to competitive classification accuracies in the incremental active phase. Yet, such deeper networks utilized more labeled and unlabeled samples in both stages to reach the maximum classification accuracy rates, in contrast to our proposed approach.

Due to the fact that our proposed approach requires limited labeled samples to train the model, it has the potential applications in recognizing emerging emotions spontaneously. In addition, our approach substantially reduces the inaccurate human annotation for complex emotion recognition. This helps in obtaining more accurate recognition of complex emotions in systems, such as emotional robots and human behavior analysis. For example, recognizing complex emotional cues helps emotion sentient robots to

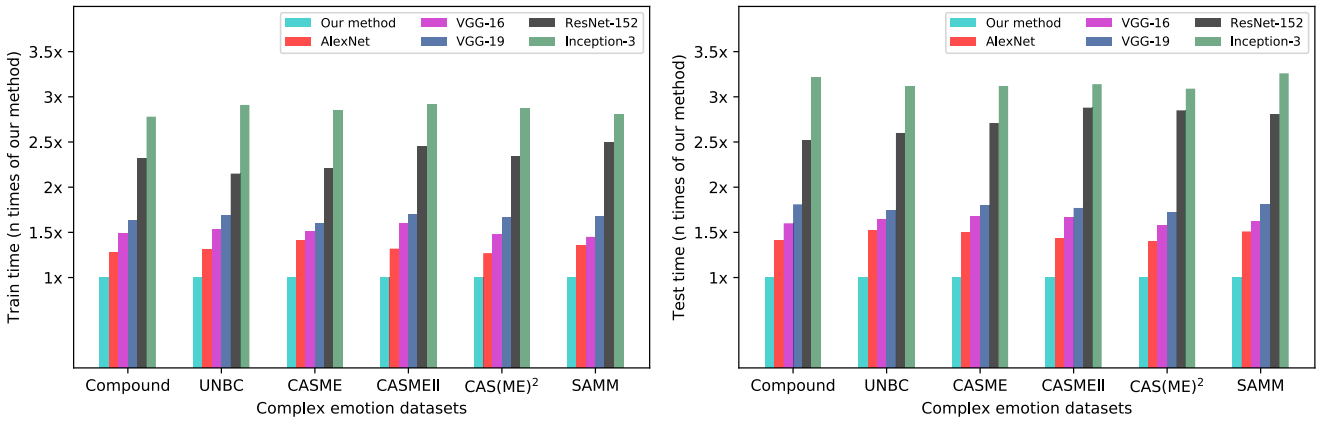

FIGURE 12. **Comparison of the computational complexity: (left) training time (right) testing time.**

respond effectively to human behaviors, and hence enhances the social interaction between the human and machines. Further, intelligent personal assistants, such as Apple's Siri, Google Assistant, Amazon's Alexa and Microsoft's Cortana can be improved to recognize the complex emotions, using our technique. In the future, they can provide personalized assistance to individuals, such as elderly people, which might help improve loneliness problems faced by the elderly. Our approach can also be improved to profile the complex emotions using videos along with audios to provide personalize assistance.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel incremental active learning-based end-to-end deep CNN framework to perform complex emotion recognition using facial expressions effectively. To the best of our knowledge, the proposed approach is the first one that exploits the use of an automatic incremental and active learning technique, to predict the complex emotions using a sparse training data accurately. Besides the key contributions of our approach, an additional advantage of our method is that there is no requirement for manual annotations during the active learning-based training process. The extensive experiments on benchmark complex emotion datasets shown that our proposed framework outperformed existing state-of-the-art deep networks and current benchmark complex emotion recognition methods. In the future, we aim to incrementally learn new complex emotions using active learning based approaches in in-the-wild environments. In addition, temporal, voice and textual features may also be considered to predict the complex emotions accurately.

REFERENCES

[1] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Emotion intensity estimation from video frames using deep hybrid convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–10.

[2] T. T. D. Pham, S. Kim, Y. Lu, S.-W. Jung, and C.-S. Won, "Facial action units-based image retrieval for facial expression recognition," *IEEE Access*, vol. 7, pp. 5200–5207, 2019.

[3] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," 2019, *arXiv:1902.01019*. [Online]. Available: http://arxiv.org/abs/1902.01019

[4] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[5] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, Oct. 2008.

[6] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5562–5570.

[7] D. Wang, C. Otto, and A. K. Jain, "Face search at scale," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1122–1136, Jun. 2017.

[8] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in *Proc. 14th ACM Int. Conf. Multimodal Interact. - ICMI*, 2012, pp. 47–52.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[10] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[11] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2323–2334, May 2015.

[12] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1183–1192.

[13] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5160–5164.

[14] C.-A. Brust, C. Käding, and J. Denzler, "Active learning for deep object detection," 2018, *arXiv:1809.09875*. [Online]. Available: http://arxiv.org/abs/1809.09875

[15] D. Mahapatra, B. Bozorgtabar, J. P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 580–588.

[16] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *J. Amer. Stat. Assoc.*, vol. 99, no. 465, pp. 67–81, 2004.

[17] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers Psychol.*, vol. 8, p. 1745, Oct. 2017.

[18] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 667–674.

[19] M. Peng, Z. Wu, Z. Zhang, and T. Chen, "From macro to micro expression recognition: Deep learning on small datasets using transfer learning," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 657–661.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[22] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.
[23] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041.
[24] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh, "An approach for automatic pain detection through facial expression," *Procedia Comput. Sci.*, vol. 84, pp. 99–106, 2016.
[25] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3466–3474.
[26] O. Rudovic, V. Pavlovic, and M. Pantic, "Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields," in *Proc. Int. Symp. Vis. Comput.* Berlin, Germany: Springer, 2013, pp. 234–243.
[27] D. L. Martinez, O. Rudovic, and R. Picard, "Personalized automatic estimation of self-reported pain intensity from facial expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 70–79.
[28] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille, "Regularizing face verification nets for pain intensity regression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1087–1091.
[29] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 84–92.
[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[31] X. Li and Y. Guo, "Adaptive active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 859–866.
[32] C. Käding, E. Rodner, A. Freytag, and J. Denzler, "Active and continuous exploration with deep neural networks and expected model output changes," 2016, *arXiv:1612.06129*. [Online]. Available: http://arxiv.org/abs/1612.06129
[33] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i-Nieto, "Cost-effective active learning for melanoma segmentation," 2017, *arXiv:1711.09168*. [Online]. Available: http://arxiv.org/abs/1711.09168
[34] S.-J. Huang, J.-W. Zhao, and Z.-Y. Liu, "Cost-effective training of deep CNNs with active model adaptation," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1580–1588.
[35] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Detecting micro-expression intensity changes from videos based on hybrid deep CNN," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2019, pp. 387–399.
[36] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.
[37] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. New York, NY, USA: Academic, 1994, pp. 474–485.
[38] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. Document Anal. Recognit.*, 2003, pp. 1–6.
[39] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *Proc. AAAI*, vol. 5, 2005, pp. 746–751.
[40] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.* Berlin, Germany: Springer, 2001, pp. 309–318.
[41] S. Argamon-Engelson and I. Dagan, "Committee-based sample selection for probabilistic classifiers," *J. Artif. Intell. Res.*, vol. 11, pp. 335–360, Nov. 1999.
[42] E. Bochinski, G. Bacha, V. Eiselein, T. J. Walles, J. C. Nejstgaard, and T. Sikora, "Deep active learning for in situ plankton classification," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2018, pp. 5–15.
[43] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.
[44] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Proc. Face Gesture*, Mar. 2011, pp. 57–64.
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[48] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Proc. Int. Symp. Vis. Comput.* Berlin, Germany: Springer, 2012, pp. 368–377.
[49] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, and M. Pietikäinen, "Capturing correlations of local features for image representation," *Neurocomputing*, vol. 184, pp. 99–106, Apr. 2016.
[50] N. Rathee and D. Ganotra, "A novel approach for pain intensity detection based on facial feature deformations," *J. Vis. Commun. Image Represent.*, vol. 33, pp. 247–254, Nov. 2015.
[51] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, and H. Wang, "Enhanced deep learning algorithm development to detect pain intensity from facial expression images," *Expert Syst. Appl.*, vol. 149, Jul. 2020, Art. no. 113305.
[52] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep hybrid spatiotemporal networks for continuous pain intensity estimation," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2019, pp. 449–461.
[53] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1–9.
[54] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, Jan. 2016.
[55] H. Zheng, X. Geng, and Z. Yang, "A relaxed K-SVD algorithm for spontaneous micro-expression recognition," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2016, pp. 692–699.
[56] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21665–21690, Oct. 2017.
[57] S. Zhang, B. Feng, Z. Chen, and X. Huang, "Micro-expression recognition by aggregating local spatio-temporal patterns," in *Proc. Int. Conf. Multimedia Modeling* Cham, Switzerland: Springer, 2017, pp. 638–648.
[58] H. Zheng, "Micro-expression recognition based on 2D Gabor filter and sparse representation," in *Proc. J. Phys., Conf.*, vol. 787, no. 1, 2017, Art. no. 012013.
[59] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "LEARNet: Dynamic imaging network for micro expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 1618–1627, 2020.
[60] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1331–1339, Nov. 2019.
[61] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184537–184551, 2019.
[62] F. Qu, S.-J. Wang, W.-J. Yan, and X. Fu, "CAS(ME)^2: A database of spontaneous macro-expressions and micro-expressions," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2016, pp. 48–59.
[63] S. P. Teja Reddy, S. Teja Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

SELVARAJAH THUSEETHAN received the B.Sc. degree from the University of Jaffna, Sri Lanka. He is currently pursuing the Ph.D. degree with the School of Information Technology, Deakin University, Geelong, Australia. He has been a Probationary Lecturer with the Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, since 2015. His current research interests include emotion recognition, computer vision, machine learning, and pattern recognition.

SUTHARSHAN RAJASEGARAR (Member, IEEE) received the Ph.D. degree from The University of Melbourne, Melbourne, VIC, Australia, in 2009. He was a Research Fellow with the Department of Electrical and Electronic Engineering, The University of Melbourne, and a Researcher in machine learning with the National ICT Australia. He is currently a Senior Lecturer with the School of Information Technology, Deakin University, Geelong, Australia. His current research interests include anomaly/outlier detection, distributed machine learning, spatio-temporal estimations, pattern recognition, computer vision, health analytics, and wireless communication.

JOHN YEARWOOD (Member, IEEE) received the B.Sc. degree from Monash University, Australia, the M.Sc. degree from Sydney University, Australia, and the Ph.D. degree from RMIT University, Australia. In 1989, he was a Lecturer with the School of Information Technology and Mathematical Science, University of Ballarat, Australia, where he was appointed as a Professor, in 2007. He is currently a Professor and the Head with the School of Information Technology, Deakin University, Australia. He has authored over 140 refereed journals, book chapters, and conference articles. His research interest includes modern optimization theory and techniques and their applications in pattern recognition, signal processing, and decision support systems.