# $MERASTC$: Micro-expression Recognition using Effective Feature Encodings and 2D Convolutional Neural network

Puneet Gupta
Discipline of Computer Science and Engineering, IIT Indore, Indore, India
Email: puneet@iiti.ac.in

**Abstract**—Facial micro-expression (ME) can disclose genuine and concealed human feelings. It makes MEs extensively useful in real-world applications pertaining to affective computing and psychology. Unfortunately, they are induced by subtle facial movements for a short duration of time, which makes the ME recognition, a highly challenging problem even for human beings. In automatic ME recognition, the well-known features encode either incomplete or redundant information, and there is a lack of sufficient training data. The proposed method, Micro-Expression Recognition by Analysing Spatial and Temporal Characteristics, $MERASTC$ mitigates these issues for improving the ME recognition. It compactly encodes the subtle deformations using action units (AUs), landmarks, gaze, and appearance features of all the video frames while preserving most of the relevant ME information. Furthermore, it improves the efficacy by introducing a novel neutral face normalization for ME and initiating the utilization of gaze features in deep learning based ME recognition. The features are provided to the 2D convolutional neural network that jointly analyses the spatial and temporal behavior for correct ME classification. Experimental results [1] on publicly available datasets indicate that the proposed method exhibits better performance than the well-known methods.

**Index Terms**—Micro-expression Recognition, Action units, Gaze feature, Deep Learning, Spatiotemporal CNN

## 1 Introduction

Analyzing facial expression is an active research area due to its indispensable role in human communication, affective computing, and psychology [1]. Unlike facial macro-expressions observed in day-to-day life, emotions can also be perceived as micro-expressions (ME). ME are the subtle facial expressions originated by the slight stretching or contraction of facial arteries located at face sub-regions like lips, eyes, and cheeks. They are originated by human reflexive behaviour, hence they are difficult to conceal and reveal the true feelings (or genuine expression) of humans [2]. Understanding the true feeling of humans is beneficial in: i) lie detection, which is further useful to avoid frauds [1]; ii) revealing latent emotions required for affective computing as in video summarization [3] and commercial advertisement rating [4]; iii) monitoring suspicious intent for psychotherapy [2]; iv) augmented reality by improving face synthesis; and v) clinical diagnosis [5]. Such wide applicability of ME has attracted the attention of the research community. But unfortunately, the performance of ME recognition is unsatisfactory even by human experts who can accurately recognize the facial macro-expressions. It is because human eyes are unable to process the ME originated for a short duration (typically 1/25 to 1/5 of a second), and that too from subtle facial movements [6]. These factors provide the motivation to propose the automatic ME recognition method in this paper.

ME recognition consists of: i) feature extraction, where relevant information related to the subtle facial movements is extracted; and ii) ME classification, where the extracted features are used to determine the ME. The feature should be extracted such that it includes both the spatial and temporal variations induced by ME. This observation has been extensively used in the literature to propose several handcrafted features like LBP-TOP [7] and spatiotemporal completed local quantization patterns (STCLQP) [8]. Such methods encode the excess amount of redundant information, making these methods computationally expensive and less discriminatory. The redundant information is reduced in some methods like main directional mean optical-flow (MDMO) [9], bi-weighted oriented optical flow (Bi-WOOF) [10] and STRCN-G [11] to improve the performance. They consider only the optical flow between the apex and offset frame. Hence their efficacies are undermined because most of the relevant information in the remaining frames is avoided. Like any other machine learning based system, these features are provided to classifiers for the ME classification. Recently, several deep learning based classification networks demonstrating promising ME classification performance have been proposed. Even though deep learning networks provide accurate results in several similar problems, their performance is restricted for ME classification due to insufficient training data.

This paper proposes a novel method, Micro-Expression Recognition by Analysing Spatial and Temporal Characteristics ($MERASTC$), which performs automatic ME recog-

[1] Implementation:
https://github.com/PuneetDurvik/Micro-expression-Recognition/

nition. It mitigates the problem of limited training data and utilization of irrelevant and redundant feature encodings to provide better ME recognition than the state-of-the-art methods. This paper provides the following contributions in the realm of ME recognition:

1) We modify and utilize the action units (AUs), landmarks, gaze, and appearance features from all the video frames. Compared to the existing feature encodings, the proposed encodings are compact and simultaneously preserve most of the relevant information about MEs. It allows us to perform better ME recognition than existing well-known methods. This paper also introduces a novel neutral face normalization for improving the efficacy of AUs in ME recognition. Furthermore, it initiates the utilization of gaze features in deep learning based ME recognition.
2) Despite the availability of small-sized datasets, the proposed 2D-CNN network can classify the ME correctly. It is possible because: the proposed network jointly analyses the spatial and temporal information; relevant data augmentation techniques improve the training; the proposed network is able to incorporate the relevant information regarding the number of video frames; and the proposed features are compact and informative.

The paper is organized in the following manner. The related work required for a better understanding of the proposed method is discussed in the next section. The proposed method, $MERASTC$ is presented in Section 3. The experimental results are analyzed in Section 4, and the conclusions are given in the last section.

## 2 Literature Survey

### 2.1 Action Units

Facial movements can be encoded using the Facial Action Coding System (FACS), which comprises 57 elementary unique structures, known as action units (AUs). These AUs are the relevant abstraction of facial expressions in terms of facial movements. Hence, they are extensively used for face macro-expressions, but in contrast, they are not well studied in the context of ME recognition. One crucial factor for this negligence is the lack of knowledge of a neutral face, which is sometimes essential for estimating the accurate facial expression [12]. The importance of a neutral face can be understood by the fact the neutral face of some persons looks happier or sadder than others. In such cases, facial emotion cannot be determined unless the attributes of the neutral face images are removed from the AUs [13]. Another crucial factor that limits the applicability of AUs in videos is the temporal scaling [14]. It arises when different ME videos contain different number of video frames. AU encoded features should be defined such that they mitigate the temporal scaling while preserving the subtle facial movements. Based on AUs, a face video can be divided into the following parts [14]: (i) onset, where AU intensities increases; (ii) apex, where AU intensities are at their peak; (iii) offset, where AU intensities decrease; and (iv) neutral, where AUs are least activated.

### 2.2 Face Micro-expression Recognition

ME analysis consists of ME spotting and ME recognition. Initially, the frames containing ME expressions are determined in ME spotting, and subsequently, they are used for ME recognition. The purpose of ME recognition is to estimate the emotion in the given ME video clip by analyzing its features in both the spatial and temporal domains. Facial expressions, both macro-expression and ME are estimated by utilizing facial geometry and/or texture features. Subsequently, a classifier is applied to the extracted feature for estimating the emotion. In contrast to macro-expression recognition, the performance of ME recognition can be improved by selecting those discriminating features that encode subtle facial movements [15].

#### 2.2.1 *Feature Extraction*

The most commonly utilized texture-based feature extraction in ME recognition is a local binary pattern in three orthogonal planes (LBP-TOP) feature [7], [16]. It encodes both the spatial and temporal local texture information by consolidating the LBP for the full face area in all three planes (viz., XY, YT, and XT). It is modified by including more relevant information regarding ME recognition. As an instance, STCLQP [8] utilizes additional information in terms of magnitude and orientation along with the LBP-TOP features for better ME recognition. Similarly, the performance is improved in [17] by analyzing the ME clip using multi-scale oriented phase variations using the Riesz pyramid. In such methods, an excess amount of redundant information is encoded, making these methods computationally expensive and less discriminatory. The redundant features can be reduced to improve performance, as in STRCN-A [11], where the redundant information from several face areas (like, chin) is removed for better ME recognition.

Apart from the texture based feature encoding, ME recognition can also be performed by utilizing facial geometry based feature encoding. This encoding requires the temporal movements or optical flow fields of facial landmarks or regions. In MDMO [9], histograms obtained from the optical flow vectors are used for ME recognition. Similarly, [18] employed optical flow information between the subsequent video frames. It requires extensive redundant information, which can be reduced to improve the ME recognition [19]. Based on this observation, Bi-WOOF [10] and STRCN-G [11] are proposed. They consider only the optical flow between the apex and offset frame to provide state-of-the-art performance. However, the efficacies of these methods are undermined because they neglect the relevant information present in the remaining frames. The efficacy of ME recognition can be improved by incorporating additional information about face macro-expressions [20], [21] and speech [22].

#### 2.2.2 *Classification*

In the ME classification, the encoded features are provided to classifiers like SVM, random forest, and deep learning based classification networks [23]. Recently, deep learning based classifiers demonstrate promising results in ME recognition. In [24], a two stream CNN network, named Off-ApexNet is proposed. These streams employ the horizontal and the vertical optical flow (OF) between the offset and apex frame for ME recognition. Eventually, these streams are merged to provide ME recognition. It is modified in STSTNet [25] which employs three streams. Along with the horizontal and the vertical OF as inputs in two streams (just like Off-ApexNet [24]), it uses OF strain as input to the remaining third stream. Similarly, triple stream CNN, TSCNN is employed in [26] where the first stream encodes the spatial context of only the apex frame; the second stream encodes the local spatial context obtained by upscaling and segmenting the video frames; and the third stream encodes the normalized optical flow between the onset and apex frame. It is possible to incorporate appearance feature encodings in deep learning based classification networks along with the geometric encoding, as in Dual-Stream Shallow Network, DSSN [27]. Since these networks utilise only the apex and offset frame, they neglect most of the relevant information in the remaining frames and thereby restricting their efficacies.

ME recognition can be improved by jointly analyzing both the spatial and temporal deformations. To this end, 3D-CNN [28], [29] or CNN network followed by sequence modeling network can be utilized. As an instance, Spatiotemporal Recurrent Convolution Network (STRCN) [11], consisting of CNN followed by recurrent convolutional layers (RCLs), performs the ME recognition by employing both appearance and geometric features. One major problem that undermines the efficacy of 3D-CNN or CNN network followed by a sequence modeling network, is the lack of sufficient training data. Hence, several features are discarded in STRCN to be trained with limited training data. To this end, only a few facial pixels are heuristically selected in STRCN appearance based feature encoding. The heuristics are inappropriate in some cases. For example, STRCN necessitate the utilization of eye areas, but they can be easily affected by eye blinking. Furthermore, the location of selected pixels is not provided to the network, which provides an important clue. Similarly, the geometric feature encoding of STRCN employs the OF between offset and apex frame. Thus, it neglects relevant information about subtle facial movements from most remaining frames, thereby restricting its efficacy. Just like STRCN [11], MERGCN proposed in [30] first extracts the AUs from the video using spatio-temporal networks. Eventually, the extracted AUs are combined using the graph convolutional network for ME recognition.

### 2.3 Openface based Face Analytics

The Openface algorithm is extensively used for face analytics. It localizes the 68 landmark points using [31], estimates the human gaze directions of both the eyes in x, y and z-directions using [32], and provides the intensities of the following AUs: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45 [33].

Neutral faces are person-specific, that is some neutral faces seem happier than others. Such person-specific neutral expression results in incorrect AU estimations and thereby wrong ME recognition [33]. For a better recognition, the effect of person-specific neutral expressions should be normalized (or removed) from the AUs. Furthermore, the normalization is also helpful in reducing the effect of domain adaptation and providing better generalization capability [34]. The model is trained and tested on different but related data distributions in domain adaptation. Usually, person-specific normalization is performed by detecting the neutral face in a video frame and removing its impact on all the video frames. In facial macro-expression videos, the neutral face features are given by the median of feature attributes. It utilizes the intuition that input face video contains many frames, and the neutral face is present in most of them [33]. In contrast, the face videos of ME contain a small number of frames, and the neural face is present in a few of them. Thus, the person specific normalization proposed in [33] is not useful for the ME recognition. The implementation of this algorithm is publicly available in [35].

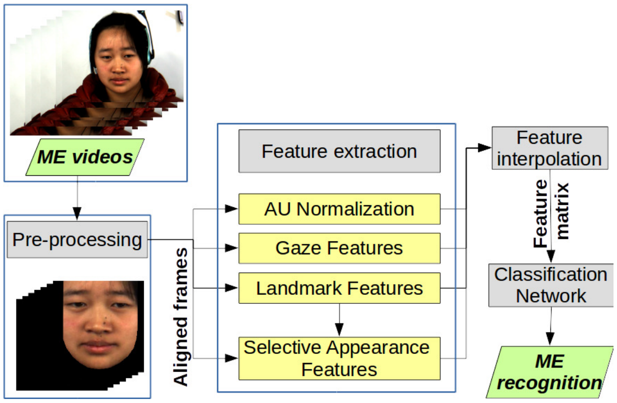


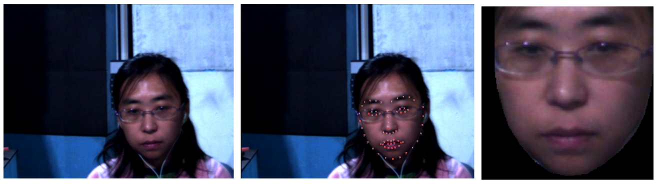
Fig. 1. Flow-graph of the proposed method, $MERASTC$. It consists of the following four stages: Preprocessing, Feature extraction, Feature interpolation, and Classification network. In preprocessing, the face is detected and aligned using landmark detection and face alignment methods proposed in [36]. Several relevant features are extracted from the face during feature extraction. For instance, the selective appearance features are given by the appearance features from cheek regions. The features are interpolated and provided to the proposed CNN based classification network (shown in Figure 6) for performing the ME recognition.

## 3 Proposed Method

In this section, the proposed ME recognition method, $MERASTC$ is presented. It consists of the following four stages: preprocessing, feature extraction, feature interpolation, and classification network. In the first stage, the face present in the video frames is aligned, and non-facial areas are removed. Several relevant features are extracted in the next stage to represent the subtle facial deformations induced by ME. In the third stage, the extracted features are interpolated to restrict their size and mitigate temporal scaling. In the last stage, the interpolated features are provided to the classification network for performing the ME recognition. The flow-graph of the proposed method, $MERASTC$ is shown in Figure 1.

### 3.1 Preprocessing

MEs are induced on the face areas, and hence the non-facial areas need to be first removed from the face videos. Furthermore, existing feature extraction methods provide erroneous ME recognition when the face present in the input video is rotated and/or translated [37]. This issue can be mitigated



a) Input frame b) Input frame with landmarks c) Aligned frame

Fig. 2. An example of Preprocessing. It depicts the input frame in a); the extracted 68 landmarks on the input frame in b); and the aligned face obtained after face normalization using both the eye center in c).

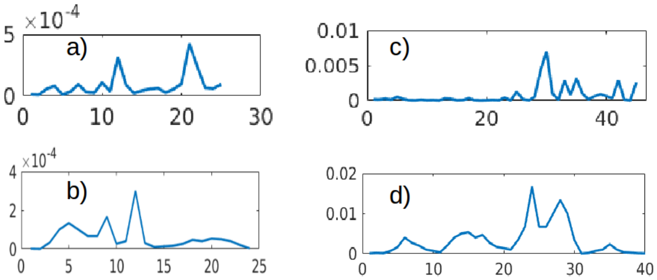


Fig. 3. Examples of Gaze features for happy ME are shown in a) and b) while gaze features for disgust ME are shown in c) and d). The x-axis of the figures represents the frame number, while the y-axis represents the gaze feature of the right eye in the x-direction. It demonstrates that there are small gaze changes for happy ME while there are higher changes for disgust ME.

when the face present in the input frames is aligned to a common reference. Both these tasks of face detection and alignment are performed in the proposed method by utilizing discriminating facial landmark points. For this purpose, the Constrained Local Neural Field model [36] is employed for detecting these discriminatory landmarks. These landmarks provide the contour of face boundary, eyes, nose, and mouth. Method [36] first applies Viola-Jones face detector [38] to detect plausible face areas, and subsequently, global and local facial models are employed to refine the facial landmark positions. The face area is correctly localized by the convex hull of the landmark points, and the remaining area is masked [39]. For face alignment, the face detected in each frame is normalized using the locations of both the eye center. The locations of eye centers are determined using the method proposed in [40]. For visualization, consider Figure 2, which shows an example of the extracted 68 landmarks and aligned face.

## 3.2 Feature Extraction

In this subsection, subtle facial deformations are extracted from the aligned face images in terms of AU, gaze, landmarks, and selective appearance features. To this end, the Openface algorithm proposed in [33] is utilized to extract the AUs, gaze, and landmarks.

### 3.2.1 AU Normalization

The extracted AUs are affected by person-specific neutral expressions, thus the AUs can be erroneous and thereby provide wrong ME recognition [33]. This issue can be mitigated by utilizing person specific normalization. Since the extensively utilized person specific normalization proposed in [33] is not useful for the ME recognition, a novel AU normalization for ME recognition is introduced in this paper.

The proposed normalization leverages the intuition that the neutral face is least deviated from their neighbouring frames. The deviation of each frame is estimated from its neighbours and selecting the frame containing least deviation as the frame containing the neural face. Assume that the aligned face images are transformed to grey-scale images, then, the frame $n$ which contains the neutral face is given by:

$$n = \arg\min_{i} \sum_{(x,y)} \sum_{p\in\{-1,1\}} \left| I_{i+p}\left(x,y\right) - I_{i}\left(x,y\right) \right| \tag{1}$$

where $I_i$ denote the grey-scale image corresponding to the $i^{th}$ video frame; $I_i\left(x,y\right)$ denotes the pixel intensity of $(x,y)$ pixel in $I_i$; the range of $(x,y)$ depends on the frame size; and $p$ is used for selecting the neighbours. The modified AU normalization is performed by subtracting the AUs of $n^{th}$ frame (that is, neutral frame) from the AUs of each frame. For clarity, assume that $\bar{A}_q^i$ denotes the $q^{th}$ normalized AUs for $i^{th}$ frame, where $q \in (1,2,4,5,6,7,9,10,12,14,15,17,20,23,25,26,45)$. Then, it is given by

$$\bar{A}_q^i = A_q^i - A_q^n \tag{2}$$

where $A_q^i$ and $A_q^n$ denote the $q^{th}$ AUs for $i^{th}$ frame and $n^{th}$ frame (containing neutral face) respectively. The size of normalized AU is 17 × $f$ where $f$ is the number of frames. Since $f$ varies depending on the ME duration, the size of normalized AU is not fixed.

### 3.2.2 Gaze Features

The human gaze provides an important clue in understanding the human emotions [41]. Positive or happy emotions are induced in humans when they perform the task that attracts them. Usually, in such cases, they dedicatedly focus on the task by nearly fixing their gaze. On the other hand, the human usually refrains from looking at those tasks that induce disgusting emotion by changing their gaze direction. Hence, the variations of gaze direction are explored in this paper for ME recognition.

The gaze extracted from [35] requires accurate eye centers. Thus, it can be erroneous due to eye blinking, which results in the closing of the eyes. This issue is mitigated by applying the particle filter [42] on the estimated gaze directions. Since three directions of both the eyes have been estimated, the size of estimated gaze features is 6 × $f$ where $f$ denotes the number of frames. Hence, the size of these features is not fixed as $f$ varies depending on the ME duration. Examples of the gaze feature are shown in Figure 3. It can be observed from the figure that there are small gaze changes for happy ME while higher changes for disgust ME.

### 3.2.3 Landmark Features

The temporal changes in landmarks locations extracted from [35] provide the relevant information about ME because they occur due to the subtle movements induced by the ME. As an instance, usually, happy emotion induces smiles, which moves the lip sideways; surprise emotion induces

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3061967, IEEE Transactions on Affective Computing

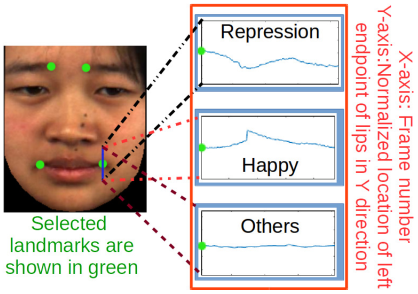
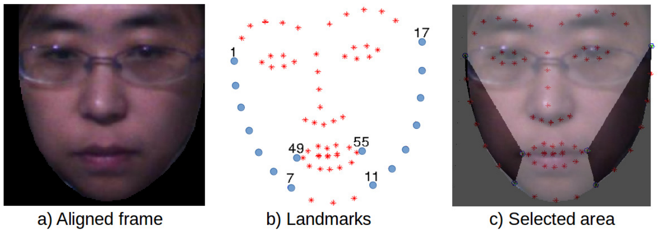


Fig. 4. Examples of the proposed landmark features. The selected landmarks are depicted on the left, and examples of encoded landmark features are depicted on the right for the following MEs: repression, happy, and others.



a) Aligned frame b) Landmarks c) Selected area

Fig. 5. Example of landmarks and appearance. An example of 68 landmarks selected from the image in a) is shown in b). Amongst them, the landmarks utilized for appearance features are shown in blue. The extracted cheek regions are shown in c) using black color.

upward movement followed by a downward movement in the eyebrows; and angry emotion induces downward movement followed by an upward movement in the eyebrows. Hence, temporal movements of distinctive landmark points of lips and eyebrows provide relevant information for ME recognition. It motivates us to incorporate the temporal movements of the endpoints of lips and both the eyebrows in the proposed ME recognition. Kindly note that the proposed method avoids using landmarks belonging to eye and face boundaries because eye areas are easily affected by eye-blinking, and the facial boundary is least affected by ME. Instead, it only utilizes the eyebrow's inner endpoints and lip endpoints. An illustrative example is shown in Figure 4, where the selected landmarks are shown in a video frame using green color.

Each landmark is defined by two 1D matrices, which contain the landmark location in x and y directions for each aligned frame. Different faces have different initial landmark locations, but only the temporal movement of landmarks are required for ME recognition. Thus, the landmark locations of each frame are normalized by subtracting the landmark locations of the first frame. Mathematically, assume that $\left(X_i^l, Y_i^l\right)$ denote the pixel location of $l^{th}$ landmark point in $i^{th}$ frame, then their normalized location $\left(\bar{X}_i^l, \bar{Y}_i^l\right)$ is given by:

$$\bar{X}_i^l = X_i^l - X_1^l \quad \text{and} \quad \bar{Y}_i^l = Y_i^l - Y_1^l \tag{3}$$

where $\left(X_1^l, Y_1^l\right)$ denote the pixel location of $l^{th}$ landmark point in the first frame. An example depicting the normalized y-locations of a lip endpoint for different ME is shown in Figure 4. It can be visualized from the figure that different ME results in different temporal movements of lip endpoints in the y-direction. Since four landmarks having two directions are utilized, the size of estimated landmark features is $8 \times f$ where $f$ denotes the number of frames. Thus, the size of these features is not fixed as $f$ varies depending on the ME duration.

### 3.2.4 *Selective Appearance Features*

Previously, most of the literature ignores the cheek areas for ME recognition [11]. Even the proposed normalized AUs, landmarks, and gaze features provide minimal importance to the cheek areas. It is recently shown in [43] that cheek areas provide valuable ME information. It motivates us to incorporate cheek appearance features in the proposed method. The cheek areas are determined by utilizing the extracted facial landmark points. That is, the left and right cheek areas are determined by determining the convex hull of the following landmarks: i) 1 to 7 and 49; and ii) 11 to 17 and 55. An example depicting the cheek areas is shown in Figure 5. Assume that the aligned face images are transformed into gray-scale images, and $f$ denotes the number of frames, then the cheek features $(C_1, C_2, ..., C_f)$ are given by:

$$C_i = \sum_{(x,y)} \left| I_i\left(x,y\right) - \frac{\sum_{j=1}^{f}\left(I_j\left(x,y\right)\right)}{f} \right| \quad for\ i = 1\ to\ f \tag{4}$$

where $I_i$ denotes the grayscale image corresponding to the $i^{th}$ video frame; $I_i\left(x,y\right)$ denotes the pixel intensity of $(x,y)$ pixel in $I_i$; and the pixels $(x,y)$ belongs to the extracted cheek regions. In essence, the cheek appearance feature in a frame is obtained by: i) selecting the pixels belonging to cheek regions; ii) normalizing their grey-scale intensities by subtracting them with their mean grey-scale intensities of all the frames; and iii) adding the normalized intensities in the frame. Thus, the size of these features is 1 × f where f denotes the number of frames, and it is dependent on the ME duration.

## 3.3 Feature Interpolation

Sequence classification deep models like RNN [11] and LSTM [43] are the most suitable for simultaneously analyzing the spatial and temporal behavior. But, it requires a large number of tunable parameters for managing the variable length sequences and long-term dependencies [44]. Similarly, other options like 3D-FCNN [29] also require a large number of tunable parameters for defining their 3D filters. A large number of parameter estimation requires a large amount of training data to achieve good ME recognition. Unfortunately, there is a scarcity in the data available for ME recognition. The number of parameters can be reduced by reducing the feature size and utilizing 2D-CNN architectures. This strategy is employed in Off-ApexNet [24] where only onset and apex frames are used for reducing the feature size. Since it neglects all the remaining frames, it avoids the complete temporal movements of distinctive facial features. This issue of neglecting most frames is eliminated in the proposed method by fixing the size of features using feature interpolation.

Each of the extracted features provides the temporal deformations in the aligned video frames. The size of each

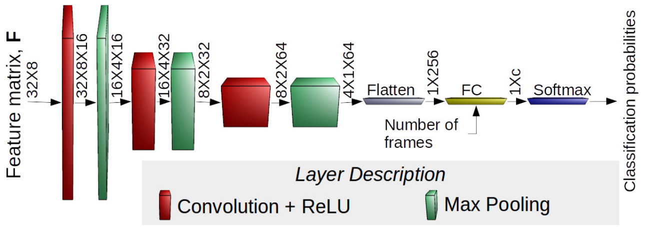


Fig. 6. Flow-graph of the proposed network architecture. Kindly note that the size of classification probabilities, $c$ depends on the number of classes but the input matrix, $\mathbf{F}$ has fixed size.

feature is restricted to a fixed value $p$ by performing 1D linear interpolation on that feature attribute. As an instance, interpolation for AU features is performed by utilizing 1D linear interpolation which interpolates the size of the normalized AU from $17 \times f$ to $17 \times p$, where $p$ is the hyper-parameter. Similarly, the interpolation of gaze, landmark, and selected appearance features is performed to restrict their size to $6 \times p$, $8 \times p$, and $1 \times p$, respectively. In this paper, the value of $p$ is set as 8. The details of this parameter selection are provided in Section 4.2. All the interpolated features are concatenated, and hence the size of the resultant feature matrix, $\mathbf{F}$ is $32 \times 8$.

## 3.4 Classification Network

This subsection initially describes the proposed deep learning based 2D CNN architecture for ME recognition. Subsequently, it describes the methodology utilized for the proper training of this network.

### 3.4.1 Network Architecture

The proposed deep learning architecture for ME recognition adopts the CNN architecture inspired by VGG-16 architecture [45]. Its flow-graph is shown in Figure 6. The feature matrix, $\mathbf{F}$ is provided as input to this network, and it provides the classification probabilities in a $c$ dimensional vector, where $c$ is the number of classes in the dataset. The ME of the input video is given by the class containing the maximum probability.

The dimension of the network input, $\mathbf{F}$ is 32×8. The first layer in this network is the convolution layer with ReLU activation consisting of 16 channels. After that, a max pooling layer is attached to reduce the computation. The combination of convolution layer with ReLU activation and max pooling layer is appended two more times in the network. The third and fifth layer are the convolution layers having 32 and 64 filters, respectively. Kindly note that each of the filters used in the convolution layer has a filter size of $3 \times 3$, and all these convolutional layers employ the same padding. Moreover, the stride of max-pooling layer is 2×2. The resultant is a 4×1×64 dimensional tensor, which is flattened to 1×256. It is observed that the number of frames plays a crucial role in ME recognition. Unfortunately, this information is absent in the interpolated features. Thus, the normalized number of frames have been introduced as an additional feature at this stage. It is given by the ratio of the number of frames and frames per second of the input video. Hence, the size of the output tensor is increased by one, and this tensor of size 1×257 is passed to the fully connected (FC) layer, and it will output a $c$ dimensional vector. Thus, the FC layer contains 257×c nodes. The result of the FL layer is further passed to the Softmax layer for normalization. Kindly note that the value of $c$, that is, the number of classes varies according to the dataset, and likewise, the proposed network varies accordingly.

### 3.4.2 Network training

The existing ME datasets contain imbalanced classes and even classes have an extremely small number of samples. It can lead to poor performance due to over-fitting. To mitigate this issue, the proposed method utilise the binary balanced loss proposed in [46] for multiple categories. It modifies the standard cross entropy loss by down-weighting the loss assigned to well-classified examples and up-weighting the loss on a sparse set of misclassified examples. Mathematically, the standard cross entropy loss, $\mathcal{L}_n$ for the $n^{th}$ sample is given by:

$$\mathcal{L}_n = \sum_{i=1}^{c} t_i^n log\left(p_i\right) \tag{5}$$

where $c$ is the number of classes; $p_i$ is the Softmax probability of $i^{th}$ class; and $t_i^n$ is the ground-truth label which is 1 when the $n^{th}$ sample belongs to the $i^{th}$ class and 0, otherwise. The modified balanced loss function, $\bar{\mathcal{L}_n}$ proposed in [46], is given by:

$$\bar{\mathcal{L}_n} = \sum_{i=1}^{c} t_i^n \alpha_i \left(1 - p_i\right)^{\gamma} log\left(p_i\right) \tag{6}$$

where $\alpha$ is set to inverse class frequency for minimizing the class imbalance problem while $(1 - p_i)^{\gamma}$ is introduced to differentiate different samples based on their classification. The parameter $\gamma$ is set to 1 in this paper.

Deep learning models require a large number of input samples for training, which are not available in ME datasets. Hence, data augmentation is utilized to increase the number of training samples. It introduces new samples by adding some deformations in the existing samples. The most commonly used augmentation technique is to apply rotation, translation, and scaling. These are not useful in the proposed ME recognition because it uses image alignment, which will eventually remove these deformations before the new samples are provided to the network. Instead, the vertical flipping and random frame pruning data augmentation techniques are employed. In vertical flipping, the image is mirrored vertically. While in random frame pruning, some percentage of frames are randomly removed from the video clip. In total, the following four levels of percentages are used for pruning the frames: 0%, 10%, 20%, and 30%. Hence, the original training data is increased by 8 (2 vertical flipping × 4 random frame pruning) times. Furthermore, the proposed network employs stochastic gradient descent (SGD) for learning the network parameters. The momentum, weight decay, and stopping criterion are set to 0.9, 0.0005, and $10^{-3}$ respectively. Initially, the learning rate is $10^{-3}$, and it will be modified in the subsequent iterations using the damping factor of 0.8. This network model is implemented in Python using the Tensorflow framework, and it uses a batch size of 100.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3061967, IEEE Transactions on Affective Computing

# 4 Experimental Results

## 4.1 Experimental Settings

The performance of the proposed method, $MERASTC$ is evaluated using the following well known publicly available spontaneous ME datasets: SMIC-HS [7], CASME II [16] and SAMM [47] and CAS(ME)$^2$ [48]. The SMIC-HS dataset constitutes 164 spontaneous ME videos, acquired from 16 subjects and recorded at 100 fps. These constitute the following MEs: positive, negative, and surprise. Likewise, the CASME II contains 247 videos acquired from 26 subjects at 200 fps. These are categorized into the following five ME classes: happiness, disgust, repression, surprise, and others. Also, the SAMM dataset contains 159 videos acquired from at 200 fps. It contains several ME classes, amongst which several classes contain a small number of ME videos. The well-known existing methods neglect those ME classes containing less than 10 ME samples. The consistency is maintained by following these guidelines, and only 136 videos of the SAMM dataset are considered, which comprise the following ME classes: anger, contempt, happiness, surprise, and others. Similarly, CAS(ME)$^2$ dataset contains 57 ME facial videos comprising the following ME classes: happy, angry, and disgust. The ground-truth ME in these datasets is provided by trained experts.

The performance of the proposed method is evaluated by utilizing the most popular methodolgy for ME recognition, which is, leave-one-subject-out (LOSO) methodology. The following two most popular testing strategies are employed for the experimentation:

1) **Strategy 1:** Here, the performance is evaluated by considering the ground-truth classes provided in the dataset. The performance metrics for this testing strategy are accuracy and $F_1$-$score$. The accuracy, $Acc$ is given by the ratio of the total number of true positives and total number of test samples while $F_1$-$score$ is given as the average of $F_1$-$scores$ of individual classes. That is,

$$F_1\text{-}score = \frac{\sum_{i=1}^{c} F_i}{c} \tag{7}$$

where $c$ denotes the number of classes and $F_i$ is the $F_1$-$score$ of $i^{th}$ class. The $F_i$ is given by the harmonic mean between precision, $P_i$ and recall, $R_i$ of $i^{th}$ class, that is,

$$F_1\text{-}score = 2 \times \frac{P_i \times R_i}{(P_i + R_i)} \tag{8}$$

for

$$P_i = \frac{TP_i}{(TP_i + FP_i)}\ ;\ R_i = \frac{TP_i}{(TP_i + FN_i)} \tag{9}$$

where $TP_i$, $FP_i$ and $FN_i$ denote the true positive, false positive and false negative for the $i^{th}$ class, respectively.

2) **Strategy 2:** In this strategy, SAMM, SMIC-HS, and CASME II datasets are merged, and the efficacy of the merged dataset is analyzed [49]. This merged dataset is referred to as $FULL$. It mimics a more realistic scenario by reducing the database bias and increasing the number of video samples for each class that facilitates deep learning techniques. The different classes of SAMM, SMIC-HS, and CASME II datasets are mapped to a common set of ME classes to perform the merging. Hence, Repression, Anger, Sadness, Contempt, Fear, and Disgust are mapped to the Negative class; Happiness is mapped to the Positive class; and Surprise remains as it is. The ME videos belonging to the remaining emotions are not considered in the $FULL$ dataset. Mainly, the ME videos belonging to Others ME class of CASME II are not considered in the $FULL$ dataset. A detailed description of this strategy is provided in [49]. The performance metrics for this testing strategy are $F_1$-$score$ and unweighted average recall, $UAR$. The $F_1$-$score$ is obtained using Equation 7. This strategy replaces $Acc$ with $UAR$ because the $FULL$ dataset contains imbalanced class distribution. Mathematically, $UAR$ is given by:

$$UAR = \frac{\sum_{i=1}^{c} TP_i}{c} \tag{10}$$

where $TP_i$ and $c$ denote the true positive for the $i^{th}$ class and number of classes, respectively.

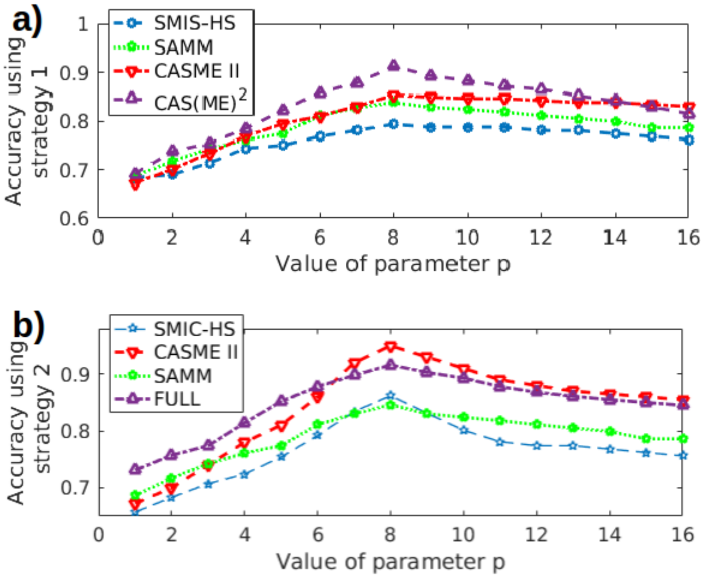


Fig. 7. Relationship between accuracy and hyperparameter $p$ for different datasets. The accuracy using strategy 1 and strategy 2 are shown in a) and b), respectively. It depicts that the best performance is achieved when $p$ is set to 8 for all the datasets in any strategy.

## 4.2 Parameter Selection

The proposed method requires the hyperparameter selection of one parameter, $p$. It is used to interpolate the extracted features in the temporal direction; that is, $p$ is required to encode the facial temporal deformations. It will be difficult to encode these temporal movements if $p$ is set to a small value. In contrast, if $p$ is set to a large value, redundant information will be encoded, which will degrade the classification performance. For proper hyperparameter selection, the performances of all the datasets are evaluated for different values of $p$. The accuracy obtained using both the strategies are shown in Figure 7a) and Figure 7b), respectively. LOSO setting is employed in these experiments. It can be observed the best performance is achieved when $p$ is set to 8 for all the datasets in any strategy.

TABLE 1
Comparative Performance of ME Recognition using strategy 1

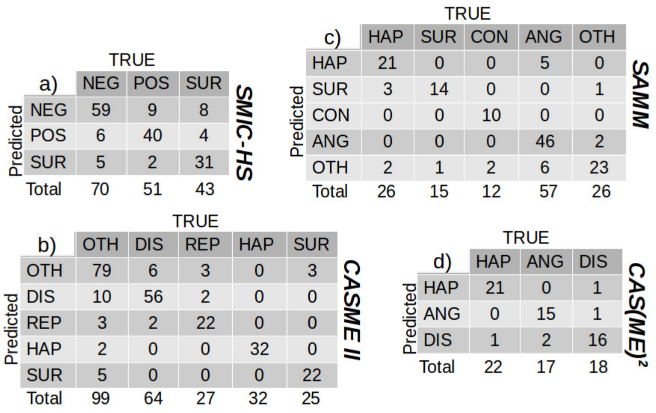
| Method | SMIC-HS | | CASME II | | SAMM | | CAS(ME)$^2$ | |
|---|---|---|---|---|---|---|---|---|
| | $Acc$ | $F_1$ Score | $Acc$ | $F_1$ Score | $Acc$ | $F_1$ Score | $Acc$ | $F_1$ Score |
| LBP-TOP+SVM [16] | 0.434 | 0.342 | 0.397 | 0.359 | 0.421 | 0.414 | 0.413 | 0.474 |
| LBP-SIP+SVM [50] | 0.445 | 0.449 | 0.466 | 0.448 | 0.423 | 0.418 | 0.419 | 0.491 |
| FDM [18] | 0.524 | 0.540 | 0.393 | 0.356 | 0.547 | 0.364 | 0.487 | 0.544 |
| BI-WOOF [10] | 0.622 | 0.620 | 0.579 | 0.610 | 0.597 | 0.605 | 0.478 | 0.596 |
| MORF + SVM [17] | 0.640 | 0.647 | 0.623 | 0.630 | 0.609 | 0.593 | 0.563 | 0.614 |
| STRCN-G [11] | 0.702 | 0.668 | 0.781 | 0.727 | 0.763 | 0.714 | 0.842 | 0.812 |
| TSCNN [26] | 0.727 | 0.723 | 0.810 | 0.807 | 0.718 | 0.694 | 0.789 | 0.772 |
| Off-ApexNet [24] | 0.682 | 0.669 | 0.761 | 0.752 | 0.632 | 0.628 | 0.737 | 0.726 |
| DSSN [27] | 0.634 | 0.642 | 0.708 | 0.730 | 0.573 | 0.464 | 0.684 | 0.662 |
| 3D-FCNN [29] | 0.555 | 0.536 | 0.587 | 0.568 | 0.559 | 0.540 | 0.439 | 0.428 |
| MER-GCN [30]$^+$ | NA | NA | 0.427 | NA | NA | NA | NA | NA |
| $MERASTC$ * | **0.793** | **0.790** | **0.854** | **0.862** | **0.838** | **0.844** | **0.912** | **0.907** |

NA: Not available in the literature.
*: $MERASTC$ is the proposed method.
$^+$: Quoted from the respective paper.

TABLE 2
Comparative Performance of ME Recognition using strategy 2

| Method | SMIC-HS | | CASME II | | SAMM | | $FULL$ | |
|---|---|---|---|---|---|---|---|---|
| | $F_1$ Score | $UAR$ | $F_1$ Score | $UAR$ | $F_1$ Score | $UAR$ | $F_1$ Score | $UAR$ |
| LBP-TOP [16] | 0.200 | 0.528 | 0.703 | 0.743 | 0.395 | 0.410 | 0.588 | 0.578 |
| Bi-WOOF [10] | 0.573 | 0.583 | 0.780 | 0.803 | 0.521 | 0.514 | 0.630 | 0.623 |
| OFF-ApexNet [24] | 0.682 | 0.669 | 0.876 | 0.868 | 0.541 | 0.539 | 0.720 | 0.710 |
| Capsule-Net [51] | 0.582 | 0.588 | 0.707 | 0.702 | 0.621 | 0.599 | 0.652 | 0.651 |
| DualNet [52] | 0.664 | 0.673 | 0.862 | 0.856 | 0.587 | 0.586 | 0.732 | 0.728 |
| STSTNet [25] | 0.680 | 0.701 | 0.838 | 0.869 | 0.659 | 0.681 | 0.735 | 0.760 |
| EMRNet [53] | 0.746 | 0.753 | 0.829 | 0.821 | 0.775 | 0.715 | 0.788 | 0.782 |
| $MERASTC$* | **0.790** | **0.862** | **0.933** | **0.950** | **0.830** | **0.846** | **0.920** | **0.916** |

*: $MERASTC$ is the proposed method.



a) SMIC-HS

| Predicted \ TRUE | NEG | POS | SUR |
|---|---|---|---|
| NEG | 59 | 9 | 8 |
| POS | 6 | 40 | 4 |
| SUR | 5 | 2 | 31 |
| Total | 70 | 51 | 43 |

c) SAMM

| Predicted \ TRUE | HAP | SUR | CON | ANG | OTH |
|---|---|---|---|---|---|
| HAP | 21 | 0 | 0 | 5 | 0 |
| SUR | 3 | 14 | 0 | 0 | 1 |
| CON | 0 | 0 | 10 | 0 | 0 |
| ANG | 0 | 0 | 0 | 46 | 2 |
| OTH | 2 | 1 | 2 | 6 | 23 |
| Total | 26 | 15 | 12 | 57 | 26 |

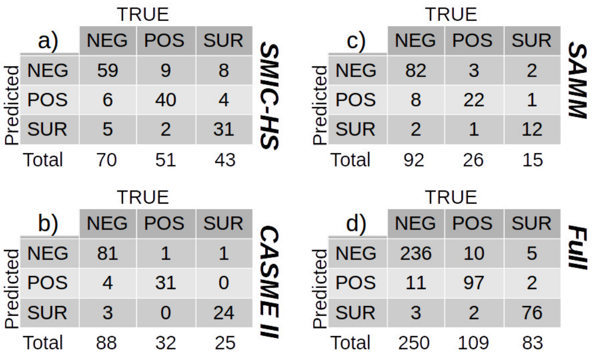
b) CASME II

| Predicted \ TRUE | OTH | DIS | REP | HAP | SUR |
|---|---|---|---|---|---|
| OTH | 79 | 6 | 3 | 0 | 3 |
| DIS | 10 | 56 | 2 | 0 | 0 |
| REP | 3 | 2 | 22 | 0 | 0 |
| HAP | 2 | 0 | 0 | 32 | 0 |
| SUR | 5 | 0 | 0 | 0 | 22 |
| Total | 99 | 64 | 27 | 32 | 25 |

d) CAS(ME)$^2$

| Predicted \ TRUE | HAP | ANG | DIS |
|---|---|---|---|
| HAP | 21 | 0 | 1 |
| ANG | 0 | 15 | 1 |
| DIS | 1 | 2 | 16 |
| Total | 22 | 17 | 18 |

Fig. 8. Confusion matrices of the proposed method $MERASTC$ on different datasets using Strategy 1. The subfigures a), b), c) and d) depicts the confusion matrix for SMIC, CASME II, SAMM, and CAS(ME)$^2$, respectively, for the following ME expressions: Happiness (HAP), Disgust (DIS), Repression (REP), Surprise (SUR), Positive (POS), Negative (NEG), Contempt (CON), Anger (ANG), and Others (OTH).



a) SMIC-HS

| Predicted \ TRUE | NEG | POS | SUR |
|---|---|---|---|
| NEG | 59 | 9 | 8 |
| POS | 6 | 40 | 4 |
| SUR | 5 | 2 | 31 |
| Total | 70 | 51 | 43 |

c) SAMM

| Predicted \ TRUE | NEG | POS | SUR |
|---|---|---|---|
| NEG | 82 | 3 | 2 |
| POS | 8 | 22 | 1 |
| SUR | 2 | 1 | 12 |
| Total | 92 | 26 | 15 |

b) CASME II

| Predicted \ TRUE | NEG | POS | SUR |
|---|---|---|---|
| NEG | 81 | 1 | 1 |
| POS | 4 | 31 | 0 |
| SUR | 3 | 0 | 24 |
| Total | 88 | 32 | 25 |

d) Full

| Predicted \ TRUE | NEG | POS | SUR |
|---|---|---|---|
| NEG | 236 | 10 | 5 |
| POS | 11 | 97 | 2 |
| SUR | 3 | 2 | 76 |
| Total | 250 | 109 | 83 |

Fig. 9. Confusion matrices of the proposed method $MERASTC$ on different datasets using strategy 2. The subfigures a), b), c), and d) depict the confusion matrix for SMIC, CASME II, SAMM, and Full dataset, respectively, for the following ME expressions: Positive (POS), Negative (NEG), and Surprise (SUR).

### 4.3 Comparative Performance Analysis

This subsection provides the comparative performance analysis between the proposed method, $MERASTC$ and the state-of-the-art methods. The performance metrics evaluated by employing strategies 1 and 2 are shown in Table 1 and Table 2, respectively. The proposed $MERASTC$ is compared with the following state-of-the-art methods: i) appearance based methods: LBP-TOP + SVM [16], and LBP-SIP + SVM [50]; ii) geometric features based methods: FDM [18], BI-WOOF [10], and MORF + SVM [17]; and iii) deep learning methods: STRCN-G + RCN [11], TSCNN [26], Off-ApexNet [24], DSSN [27], 3D-FCNN [29], and MER-GCN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3061967, IEEE Transactions on Affective Computing

[30]. Kindly note that the following two ME recognition methods are proposed in [11]: STRCN-A and STRCN-G which utilise the appearance and geometric features respectively. They observe that STRCN-G provides better recognition for LOSO protocol hence, the results are compared only with STRCN-G.

Table 1 indicates that geometric features based methods can outperform existing well known deep learning based methods. As an instance, BI-WOOF [10], and MORF + SVM [17] perform better than 3D-FCNN [29]. It is because the 3D-FCNN [29] requires a large number of redundant features derived from all the frames. In contrast, the Off-ApexNet [24], which uses a similar network, performs much better than 3D-FCNN because it avoids a significant amount of redundant information by considering only the onset and apex frame. Off-ApexNet also removes the relevant temporal information required for ME recognition as a downside. Fortunately, the proposed $MERASTC$ utilizes all the frame information and simultaneously encodes the most relevant information compactly. Due to these capabilities, it can be observed from Table 1 that the proposed method, $MERASTC$ outperforms these well-known methods.

The CAS(ME)$^2$ contains a small number of ME face videos, which are 57. Moreover, it contains both the ME and macro-expressions. Due to these reasons, it is not extensively used in the literature, and only a few methods have been tested on CAS(ME)$^2$ dataset. The well-known method for CAS(ME)$^2$ dataset is TSCNN [26] which works on a subset of CAS(ME)$^2$ dataset and achieves an $Acc$ and $F_1$ score of 0.862 and 0.862, respectively. They have utilized only those videos which contain more than 96 frames. Thus, they utilize only 40 videos. When the proposed $MERASTC$ is applied under this same condition, it achieves an $Acc$ and $F_1$ score of 0.925 and 0.917, respectively, which is better than TSCNN [26]. Thus, this observation, along with the performance metrics of Table 1 and Table 2 indicate that the proposed $MERASTC$ outperforms the well-known methods.

Furthermore, the Table 2 demonstrates that the proposed $MERASTC$ performs better than the existing well-known methods even if these datasets are combined (that is, $FULL$ dataset). Moreover, a significant increase in performance can be observed when the combined dataset is used because it contains a large number of video samples for each class that facilitate the utilization of deep learning techniques. Kindly note that it may be apparent that the performance of $MERASTC$ on the $FULL$ dataset should decrease because it is affected by database biases in the sense that it requires the number of ME frames. However, this issue is mitigated in $MERASTC$ by normalizing it with the frames per second.

The confusion matrices obtained by applying the proposed method $MERASTC$ on SMIC-HS, CASME II, SAMM, and CAS(ME)$^2$ datasets using strategy 1, are shown in Figure 8. Furthermore, the confusion matrices obtained by applying the proposed method $MERASTC$ on SMIC-HS, CASME II, SAMM, and $FULL$ datasets using strategy 2, are shown in Figure 9. It indicates that all the classes have similar performances; that is, the proposed method effectively handles the problem of unbalanced classes. Moreover, the model was trained and tested on a GPU Server with NVIDIA Tesla T4 (16 GB) processor. The approximate inference time for the ME face videos containing 30 frames is about 0.473 seconds. The majority of this time is taken by Openface [35] which is 0.367 seconds. Unfortunately, this step is inevitable, and hence, just like existing methods, the proposed $MERASTC$ performs the ME recognition in near real-time as most of the ME face videos are acquired at 200 fps.

### 4.4 Ablation Study

This subsection provides a better understanding of the efficacy of the proposed feature encodings and thereby provide a more rigorous analysis of the proposed method, $MERASTC$. The proposed $MERASTC$ utilizes the following four features for ME recognition: normalized AU, gaze, landmarks, and appearance (in cheek regions). Efficacies of these features are demonstrated by comparing several methods in Table 3. These methods are designed by using different pairing of features for ME recognition. The description of these methods is provided in the table. Kindly note that the method $AUN$-$L$-$A$-$G$-$ME$ employs all the features, and it is different from the proposed method, $MERASTC$ in the sense that $MERASTC$ additionally requires the number of frames.

Table 3 demonstrates that the AU features exhibit better performance than the landmark, gaze, and cheek features (refer $AU$-$ME$, $G$-$ME$, $L$-$ME$ and $A$-$ME$). It is because landmark, gaze, and cheek features utilize the temporal movements of some small facial regions while AU features encompass the temporal movements of the full face. Furthermore, it can be observed from Table 2 and Table 3 that the ME recognition performance obtained by utilizing only AU features (that is, $AU$-$ME$) is sometimes better than some of the well-known methods. This performance can also be further improved by incorporating the proposed person specific normalization in AU features (refer, $AU$-$ME$ and $AUN$-$ME$ in Table 3). Apart from the normalized AU features, the significant recognition performance can be obtained when only the proposed gaze, landmark or appearance features are used (refer $G$-$ME$, $L$-$ME$ and $A$-$ME$ in Table 3). All these observations point out that the proposed features properly encode distinctive facial movements and can be utilized for ME recognition.

Table 3 indicates that the proposed method, $MERASTC$, and the method $AUN$-$L$-$A$-$G$-$ME$ employing all the features, exhibit better recognition performance than the other methods. It justifies the incorporation of all the features in ME recognition. Evident performance degradation can be observed from the table when any selected features are not utilized for ME recognition. Also, $MERASTC$ performs better than $AUN$-$L$-$A$-$G$-$ME$ in Table 3, which points out that the proper knowledge of the number of video frames plays a crucial role in ME recognition.

## 5 Conclusions

ME recognition plays a crucial role in many real-world scenarios. Automatic ME recognition is highly challenging because training data is insufficient, and the existing

TABLE 3
Ablation Study using strategy 1

| Method | SMIC-HS | | CASME II | | SAMM | | CAS(ME)$^2$ | | Description |
|---|---|---|---|---|---|---|---|---|---|
| | *Acc* | $F_1$ Score | *Acc* | $F_1$ Score | *Acc* | $F_1$ Score | *Acc* | $F_1$ Score | |
| $AU\text{-}ME$ | 0.570 | 0.566 | 0.599 | 0.643 | 0.625 | 0.623 | 0.666 | 0.640 | Using only AU features |
| $AUN\text{-}ME$ | 0.611 | 0.618 | 0.648 | 0.690 | 0.662 | 0.664 | 0.701 | 0.688 | Using only normalized AU features |
| $G\text{-}ME$ | 0.460 | 0.446 | 0.478 | 0.512 | 0.485 | 0.461 | 0.526 | 0.502 | Using only gaze features |
| $L\text{-}ME$ | 0.465 | 0.447 | 0.482 | 0.573 | 0.493 | 0.472 | 0.579 | 0.557 | Using only landmark features |
| $A\text{-}ME$ | 0.407 | 0.404 | 0.385 | 0.447 | 0.382 | 0.412 | 0.474 | 0.455 | Using only appearance features |
| $AUN\text{-}G\text{-}ME$ | 0.639 | 0.635 | 0.684 | 0.740 | 0.698 | 0.723 | 0.719 | 0.702 | Using normalized AU & gaze features |
| $AUN\text{-}L\text{-}ME$ | 0.686 | 0.695 | 0.737 | 0.798 | 0.735 | 0.758 | 0.771 | 0.792 | Using normalized AU & landmark features |
| $AUN\text{-}A\text{-}ME$ | 0.633 | 0.627 | 0.692 | 0.785 | 0.713 | 0.716 | 0.754 | 0.731 | Using normalized AU & appearance features |
| $G\text{-}L\text{-}ME$ | 0.506 | 0.482 | 0.525 | 0.593 | 0.544 | 0.498 | 0.631 | 0.617 | Using gaze & landmark features |
| $G\text{-}A\text{-}ME$ | 0.477 | 0.474 | 0.482 | 0.507 | 0.515 | 0.527 | 0.596 | 0.584 | Using gaze & appearance features |
| $L\text{-}A\text{-}ME$ | 0.471 | 0.458 | 0.486 | 0.527 | 0.500 | 0.512 | 0.614 | 0.593 | Using landmark & appearance features |
| $AUN\text{-}G\text{-}L\text{-}ME$ | 0.720 | 0.719 | 0.781 | 0.799 | 0.787 | 0.792 | 0.807 | 0.790 | Using normalized AU, gaze & landmark features |
| $AUN\text{-}G\text{-}A\text{-}ME$ | 0.669 | 0.662 | 0.729 | 0.752 | 0.750 | 0.749 | 0.790 | 0.778 | Using normalized AU, gaze & appearance features |
| $AUN\text{-}L\text{-}A\text{-}ME$ | 0.716 | 0.721 | 0.806 | 0.813 | 0.787 | 0.794 | 0.860 | 0.849 | Using normalized AU, landmark & appearance features |
| $L\text{-}A\text{-}G\text{-}ME$ | 0.524 | 0.512 | 0.547 | 0.567 | 0.559 | 0.522 | 0.649 | 0.631 | Using landmark, appearance & gaze features |
| $AUN\text{-}L\text{-}A\text{-}G\text{-}ME$ | 0.744 | 0.742 | 0.834 | 0.845 | 0.823 | 0.829 | 0.894 | 0.882 | Proposed but excluding mormalised number of frames |
| $MERASTC$ | **0.793** | **0.790** | **0.854** | **0.862** | **0.838** | **0.844** | **0.912** | **0.907** | **Proposed** |

features provide either incomplete or redundant information about the subtle facial deformations induced by ME. This paper has presented a novel ME recognition method, $MERASTC$, where the most relevant features are compactly represented and provided to a deep learning network. It considered AUs, landmarks, gaze, and appearance features of all the video frames along with the normalized number of frames for the recognition. The efficacies of these features have been improved and eventually consolidated to achieve better ME recognition performance than the existing methods. The proposed method has introduced a neutral face normalization for modifying the AUs features to perform better ME recognition. Likewise, it has initiated the utilization of gaze features in deep learning based ME recognition. The features have been provided to the proposed novel network architecture for the consolidation and joint analysis of the spatial and temporal behavior. The rigorous experiments conducted on three publicly available ME recognition datasets have demonstrated that the proposed method $MERASTC$ exhibited better performance than the well-known existing methods. In the future, the proposed method $MERASTC$ will be improved so that it can be used for cross-dataset ME recognition where training and testing are performed on different datasets.

## References

[1] P. Ekman, "Lie catching and microexpressions," *The philosophy of deception*, pp. 118–133, 2009.

[2] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of research in psychotherapy*. Springer, 1966, pp. 154–165.

[3] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *ACM international conference on image and video retrieval (CIVR)*. ACM, 2009, p. 31.

[4] P. Lewinski, M. L. Fransen, and E. S. Tan, "Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli." *Journal of Neuroscience, Psychology, and Economics*, vol. 7, no. 1, p. 1, 2014.

[5] T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," *British journal of clinical psychology*, vol. 45, no. 4, pp. 579–583, 2006.

[6] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan, and J. See, "Eulerian emotion magnification for subtle expression recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1243–1247.

[7] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[8] 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, 2016.

[9] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2015.

[10] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[11] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2019.

[12] D. Neth and A. M. Martinez, "Emotion perception in emotionless face images suggests a norm-based representation," *Journal of vision*, vol. 9, no. 1, pp. 5–5, 2009.

[13] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[14] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*. IEEE, 2006, pp. 149–149.

[15] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing: Image Communication*, vol. 47, pp. 170–182, 2016.

[16] 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, p. e86041, 2014.

[17] C. A. Duque, O. Alata, R. Emonet, H. Konik, and A.-C. Legrand, "Mean oriented riesz features for micro expression classification," *Pattern Recognition Letters*, vol. 135, pp. 382–389, 2020.

[18] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[19] X. Ben, P. Zhang, R. Yan, M. Yang, and G. Ge, "Gait recognition and micro-expression recognition based on maximum margin

projection with tensor representation," *Neural Computing and Applications*, vol. 27, no. 8, pp. 2629–2646, 2016.

[20] X. Jia, X. Ben, H. Yuan, K. Kpalma, and W. Meng, "Macro-to-micro transformation model for micro-expression recognition," *Journal of Computational Science*, vol. 25, pp. 289–297, 2018.

[21] X. Ben, X. Jia, R. Yan, X. Zhang, and W. Meng, "Learning effective binary descriptors for micro-expression recognition transferred by macro-information," *Pattern Recognition Letters*, vol. 107, pp. 50–58, 2018.

[22] X. Zhu, X. Ben, S. Liu, R. Yan, and W. Meng, "Coupled source domain targetized with updating tag vectors for micro-expression recognition," *Multimedia Tools and Applications*, vol. 77, no. 3, pp. 3105–3124, 2018.

[23] K. M. Goh, C. H. Ng, L. L. Lim, and U. Sheikh, "Micro-expression recognition: an updated review of current trends, challenges and solutions," *The Visual Computer*, vol. 36, no. 3, pp. 445–468, 2020.

[24] Y. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "Off-apexnet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129–139, 2019.

[25] S.-T. Liong, Y. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–5.

[26] 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. 184 537–184 551, 2019.

[27] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 36–40.

[28] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[29] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Analysis and Applications*, vol. 22, no. 4, pp. 1331–1339, 2019.

[30] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "MER-GCN: Micro expression recognition based on relation modeling with graph convolutional network," *arXiv preprint arXiv:2004.08915*, 2020.

[31] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for 3d facial landmark detection," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2519–2528.

[32] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3756–3764.

[33] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–6.

[34] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan, "Local relationship learning with person-specific shape regularization for facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 917–11 926.

[35] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2018, pp. 59–66.

[36] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2610–2617.

[37] P. Gupta, B. Bhowmick, and A. Pal, "Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1316–1323.

[38] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[39] P. Gupta, B. Bhowmick, and A. Pal, "MOMBAT: Heart rate monitoring from face video using pulse modeling and bayesian tracking," *Computers in Biology and Medicine*, p. 103813, 2020.

[40] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[41] K. N'diaye, D. Sander, and P. Vuilleumier, "Self-relevance processing in the human amygdala: gaze direction, facial expression, and emotion intensity." *Emotion*, vol. 9, no. 6, p. 798, 2009.

[42] J. H. Kotecha and P. M. Djuric, "Gaussian sum particle filtering," *IEEE Transactions on signal processing*, vol. 51, no. 10, pp. 2602–2612, 2003.

[43] H. Shahar and H. Hel-Or, "Micro expression classification using facial color and deep learning methods," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 0–0.

[44] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE international conference on computer vision (ICCV)*, 2017, pp. 2980–2988.

[47] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2016.

[48] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)$^2$: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 424–436, 2017.

[49] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019-the second facial micro-expressions grand challenge," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–5.

[50] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2014, pp. 525–537.

[51] N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–7.

[52] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–5.

[53] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *IEEE international conference on automatic face & gesture recognition (FG)*. IEEE, 2019, pp. 1–4.

**Puneet Gupta** Dr. Puneet Gupta received his Doctoral degree from the Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, India in 2016. Currently, he is working in Indian Institute of Technology, Indore as Assistant Professor. Prior to that, he was a member of Machine Vision group in Embedded Methods and Robotics, TCS Research & Innovation. His area of research includes Biometrics, Image Processing, Computer Vision and Machine Learning. He has published several papers in the reputed International Journals and International Conferences.