Micro-expression recognition from local facial regions

M. Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, Renaud Seguier

Micro-Expression Recognition from Local Facial Regions

Mouath Aouayeb[a,b], Wassim Hamidouche[a], Catherine Soladie[b], Kidiyo Kpalma[a] and Renaud Seguier[b]

[a]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, F-35000, France
[b]Univ. Rennes, CentraleSupélec, CNRS, IETR - UMR 6164, Rennes, F-35000, France

ARTICLE INFO

Keywords:
Micro-Expression
Regions of Interest
Convolutional Neural Network
Long Short Term Memory

ABSTRACT

MiE is a facial involuntary reaction that reflects the real emotion and thoughts of a human being. It is very difficult for a normal human to detect a Micro-Expression (MiE), since it is a very fast and local face reaction with low intensity. As a consequence, it is a challenging task for researchers to build an automatic system for MiE recognition. Previous works for MiE recognition have attempted to use the whole face, yet a facial MiE appears in a small region of the face, which makes the extraction of relevant features a hard task. In this paper, we propose a novel deep learning approach that leverages the locality aspect of MiEs by learning spatio-temporal features from local facial regions using a composite architecture of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). The proposed solution succeeds to extract relevant local features for MiEs recognition. Experimental results on benchmark datasets demonstrate the highest recognition accuracy of our solution with respect to state-of-the-art methods.

1. Introduction

Human face analysis has attracted a wide interest of the computer vision research community with a potential scope of applications including healthcare, education, security, etc. The human face holds essential information for a machine to understand human reactions and feelings, and also for humans to better understand each other in interpersonal communication. Recently, particular attention is paid to a special type of expression called facial *micro-expression* (MiE) which may help to understand the feelings of a person.

The pioneering work on human facial expressions was published by Charles Darwin and Phillip Prodger [6] in 1872 with a hypothesis that some facial expressions may appear to reflect the real emotion when someone tries to hide it. In 1966, Haggard and Isaacs [20] first discovered the phenomenon of MiE when studying films of communication between a therapist and a patient frame by frame, searching for non-verbal reactions. Later in 1969, Ekman *et al.* [15] proved the existence of MiEs as quick, universal, and spontaneous facial expressions of some local regions of the face.

The analysis of MiE can be used for several applications like lie detection, pain or stress detection, and teaching assistance. MiEs, like any other facial expression, is a representation of facial muscles movements. Yet, they are hard to analyze since they are local and lasting only from 1/25s to 1/5s according to the study conducted in [14]. A precision of 40% in MiE recognition was reported in [12] for people who had passed professional training. To help reading these movements and encoding MiEs, a Facial Action Coding System (FACS) is used by P. Ekman based on the combination of Action Units (AUs). An AU is a small region of the face, controlled by one muscle or a group of muscles that react together producing a basic movement in the face. The FACS system was originally developed by Hjortsjö [22] in 1969 and then adopted in 1978 by Ekman *et al.* [13]. Figure 1 illustrates the FACS as defined by Ekman *et al.* in [16]. Ekman has also shown in [11] that human emotions can be categorized into 7 basic classes including disgust, surprise, happiness, fear, anger, contempt, and sadness.

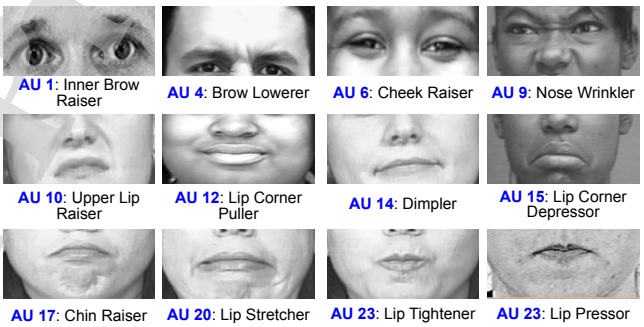

Figure 1: Face and some Action Unit locations according to the Facial Action coding System proposed by Ekman [16]. Images from http://www.cs.cmu.edu/~face/facs.htm.

Figures 2(a) and 2(b) show examples of emotions in Macro-Expression (MaE) and MiE and their corresponding AU's code according to the FACS system. The deformation of AUs defines the facial expression. This deformation goes through three main steps: Onset, Apex, and Offset [12]. The Onset represents the starting point of the expression, the Apex represents the maximum or the peak that the expression reaches in deformation, and the Offset is the instant when the expression vanishes away. Hence, the period of any facial expression is the time spent between the beginning of the Onset period and the end of the Offset period. Figure 3 shows an example of a deformation evolution through time (only 5 frames out of the sequence are shown[1]) for *disgust* emotion. In the Apex frame, where the peak of the deformation occurs, we can notice a clear contraction of muscles between

ORCID(s): 0000-0001-7411-4375 (M. Aouayeb)

[1]A Micro-Expression sequence labeled *disgust* in CASME II dataset [47]

(a) Examples of Macro-Expressions with their corresponding action units (AUs). The images are from Cohn-Kanade (CK) dataset.

(b) Examples of Micro-Expressions with their corresponding action units (AUs). Figures from [5].

Figure 2: Macro- and Micro-Expressions in FACS system.

eyebrows and eyes, which refers to the AU9.

Analogously to the progress of MaE recognition, which has been first tackled by handcrafted features then modern deep neural networks, MiE recognition methods have also evolved from hand-designed approaches [49, 28, 8] to deep spatiotemporal neural networks [24, 37, 43, 42] and hybrid approaches [17, 27, 45, 19]. However, while MaE recognition has seen dramatic gains in accuracy, improvements in MiE recognition have been more moderate.

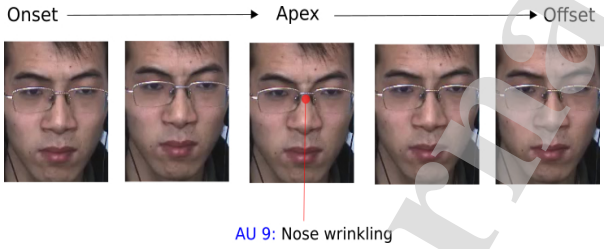

Figure 3: A sequence of 5 frames representing the evolution of a micro-expression from Onset up to Apex an then Offset.

The main drawback of previous approaches for automatic MiE recognition is the use of the whole face while a MiE occurs only in small parts of the face. These solutions have mainly contributed to the architecture of models for features extraction and classification, which mostly deal with the low intensity and short duration characteristics of MiEs and not the locality character. Nevertheless, recent works [50, 38, 51] pay attention to this locality feature. They apply either handcrafted, deep learning or hybrid approaches only on some selected regions, which may result in better performance. Since the MiEs are brief and local, and the available MiEs datasets are very small and unbalanced between classes, the result is still less accurate than the results of MaEs recognition systems.

This paper studies a new model to automatically recognize MiE from local facial regions using a composite architecture of CNN and LSTM for spatiotemporal features extraction. Our proposed solution differs from other region-based solutions since it associates a novel label for each region based on the definition of facial expressions on the FACS system illustrated in Figure 2. Also, a shallow CNN model is used instead of deep CNN, because a very deep architecture may eliminate the spatial features that are helpful for MiE analysis. The contributions of this work can be summarized as follows:

- We use a shallow CNN, combined with LSTM to treat the spatiotemporal information. By this architecture, we address the low intensity (CNN) and rapidity (LSTM) characters of facial Micro-Expression.
- The proposed architecture of CNN-LSTM is applied on 6 regions of the face to address the locality aspect.
- The training process of CNN (spatial part) is done using specialized labels for each region depending on the expression to aid the CNN learning more specific and useful features.
- We test the proposed solution on two different distribution of datasets : the Micro-Expression Grand Challenge (MEGC) 2019 challenge dataset with 3 combined data bases (CASMEII, SAMM, SMIC) and 3 classes (Positive, Negative and Neutral), and the AU-based datasets (CASMEII and SAMM) with 5 classes based on AU.

Experiments on 3 spontaneous MiE datasets demonstrate the potential of our approach performance that exceeds the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 presents the state-of-the-art solutions for MiE recognition. Section 3 investigates the proposed spatiotemporal architecture for MiE recognition. The performance of the proposed solution is assessed and compared to the best-performing solutions in Section 4. Finally, Section 5 concludes this paper.

2. Related Work

In this section, we review and discuss the state-of-the-art solutions for MiEs recognition. The overall pipeline, shown in Figure. 4, has two main parts: spatiotemporal features extraction and classification.

Given a fixed set of k classes $c_1, c_2, \ldots, c_k$ and a MiE sequence S composed of N frames $(F_1, F_2, \ldots, F_N)$, the goal of MiE recognition as a video sequence classification problem is to identify the class to which the MiE sequence belongs. In other words, we are interested in the class c_i with the highest probability $P(c_i|S)$ among the probabilities of k

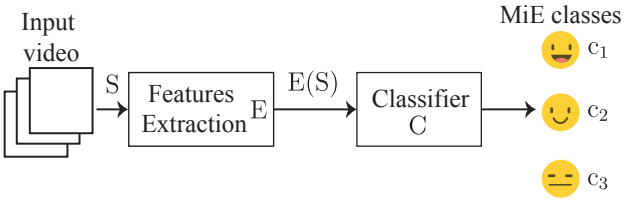

Figure 4: Pipeline of state-of-the-art micro-expression recognition task.

classes. This probability can be parameterized using different methods $f(S) = f_C(f_E(S))$ which leverage at all frames in the sequence to predict $P(c_i|S)$ as follows:

$$\begin{aligned} c_i &= argmax P(c_k|S), k \in [1, \ldots, N], \\ c_i &= argmax f_C\left(f_E(S)\right). \\ f_E(S) &= f_E(\{F_j, j = 1, \cdots, N\}). \end{aligned} \tag{1}$$

where f_E represents the model used for spatiotemporal features extraction and f_C represents the model used for the classification. Based on the proposed problem formulation, state-of-the-art works can be grouped into 4 categories including handcrafted, deep learning, hybrid, and region-based. These categories may lead to different approaches for MiE recognition.

2.1. Handcrafted approaches

The first group of solutions for MiE recognition rely on handcrafted models. The model includes spatial and temporal features. The extraction of the features is based either on appearance face information that describes the variation of pixel's intensity (texture) or on geometric face information like the shapes and the locations of facial landmarks. Zhao *et al.* [49] first introduced Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) for features extraction from dynamic textures to analyze the MiE. The features are defined in the LBP-TOP histogram which is the concatenation of the Local Binary Pattern (LBP) histograms of the three orthogonal planes: XY, XT, and YT. Concatenating histograms in the computation of LBP-TOP leads to redundant information where each neighbor pixel is used more than once. Subsequently, Wang *et al.* [44] proposed an improved version of LBP-TOP with six intersection points of the three orthogonal planes XY, XT, and YT. Moreover, LBP Mean Orthogonal Planes (LBPMOP) computes the LBP of average planes for three orthogonal planes to reduce the redundancy. Guo *et al.* [18] proposed Extended LBPTOP (ELBPTOP) that uses the second-order discriminative information in the radial and angular directions of a local path along with the normal LBP-TOP.

Chaudhry *et al.* [2] were the first to propose Histogram of Oriented Optical Flow (HOOF) for human action recognition and it was then adopted by Davison *et al.* [8] to be one of the baseline methods for MiE recognition. Main Directional Mean Optical Flow (MDMO) proposed by Liu *et al.* [30] to describe local facial dynamics by extracting principal Optical Flow (OF) direction of some AUs. Furthermore, Liong *et al.* exploited the Bi-Weighted Oriented Optical Flow (Bi-WOOF) in [28].

Different from these methods, Mean Oriented Riesz Features (MORF) [10] is proposed by Duque *et al.*. It uses the Riesz pyramid to model the temporal evolution of the MiE in two frames called Mean Oriented Riesz (MOR) image pair, which is then used to build a histogram. Polikovsky *et al.* [36] have used a 3D-Histogram of Oriented Gradient (HOG) on 12 facial regions to recognize MiE. Lu *et al.* [31] proposed the Fusion of Motion Boundary Histogram (FMBH) technique to extract features from the face. The FMBH is the combination of Motion Boundary Histograms (MBHs) which are based on the norms and the angles computed from the horizontal and the vertical components of the optical flow. More details about the MBH descriptor can be found in [4].

2.2. Deep learning approaches

Computer vision tasks like object detection and tracking, video classification, and image segmentation are widely addressed using deep learning techniques. Due to the high performance of deep learning models on these tasks, many researchers have tested and adopted deep learning architectures for MiE analysis. State-of-the-art deep learning solutions are usually based on variants of CNN or a combination of CNN and Recurrent Neural Network (RNN).

The first deep learning solution for MiE recognition was proposed by Patel *et al* [9] relying on a CNN. They used the ImageNet-VGG-f CNN [1] pre-trained on ImageNet dataset with a CNN trained on facial expressions datasets (CK+ [32] and SPOS [35]) to extract features. Since then, numerous deep learning solutions [38, 42, 24] are proposed. Reddy *et al.* [38] proposed a MicroExpSTCNN architecture that is based on 3D-CNN. Lateral Accretive Hybrid Network (LEARNet) was proposed by Verma *et al.* [42] which is based on a CNN with an accretion layer to refine the salient expression features by accretion of the learning capability of the network. One of the best derivatives of RNN is LSTM. An architecture of CNN combined with LSTM has been proposed by Kim *et al.* [24]. The CNN enables extracting the spatial features and the LSTM extracts the temporal features from the spatial features. While most computer vision tasks have seen dramatic gains in accuracy, improvements in MiE video analysis have been more moderate. This is mainly caused by the lack of MiE datasets to train deep architectures.

2.3. Hybrid approaches

Since previous categories don't achieve expected performance, many hybrid systems have been proposed. Hybrid solutions [17, 27, 19] are based on a combination of handcrafted and deep learning methods. A typical method of Optical Flow (OF) and its variants are usually employed with CNN for features extraction. Liong *et al.* [17] proposed Off-ApexNet as a hybrid solution. They consider only two frames to represent the MiE: Onset and Apex. Then, for feature extraction, they compute OF from the two considered frames

and pass the computed features to a CNN. Later, Shallow Triple Stream Three-dimensional CNN (STSTNet) [27] was proposed as an improvement of the Off-ApexNet method. The authors added to the horizontal and vertical OF, the strain of the OF. Recently, Khor *et al.* [19] proposed the Enriched Long-term Recurrent Convolutional Network (ELRCN) which can be summarized into three main steps. First, they compute the optical flow and the optical strain. Then, they suggest two different methods for learning spatial features. One method is called spatial dimension enrichment, where the spatial features Φ_S are the output of the VGG-16 [40] model trained on the concatenation of the image, the optical flow, and the optical strain. The other proposed method for spatial features extraction is called temporal dimension enrichment that trains three VGG-16 networks on the image, the optical flow, and the optical strain and then concatenates the outputs to form the spatial features Φ_T. Finally, an LSTM is applied to the extracted spatial features (Φ_S or Φ_T) for temporal learning and ends with a Fully Connected Layer (FCL) for the classification. Recently, Xia *et al.* [45] introduced Spatiotemporal Recurrent Convolution Network (STRCN) as a solution for MiE recognition. They described two versions of the solution. One version is called STRCN-A based on appearance connectivity where the image is represented in a one-dimensional vector and thus, the sequence by a 2D-matrix that will be fed to the STRCN block. The other version is STRCN-G with geometric-based connectivity where an OF is applied and the output is fed to the STRCN. The STRCN block is based on recurrent CNN.

The advantage of the methods in this category is the use of handcrafted methods to facilitate the spatiotemporal features extraction and the classification with the used deep learning architecture.

2.4. Region-based approaches

In contrast to previous cited approaches which focused on improving the feature extraction model, the contributions of region-based solutions address the locality character in the preprocessing step. Instead of using the whole face for MiE analysis, other researchers consider using some particular regions. Hence, enforce the system to extract more relevant and robust spatial features. Such region-based methods have achieved state-of-the-art performance. Ekman *et al.* [16] have already identified six different regions that are left and right (eye+eyebrow), the nose, the two cheeks, and the mouth. Based on these regions, Zhao *el al.* [50] have selected manually 18 regions of interest called Active Patchs (APs). From these APs, they identified Necessary Morphological Patches (NMPs) by giving a weight for each AP. The weights are calculated using an entropy-weight method based on their extracted features by LBP-TOP. An improved version to identify the NMPs was proposed by Zhao *et al.* [51]. Instead of choosing manually 18 APs, they divide the six regions into small blocks to get 106 APs and apply the Random Forest (RF) algorithm on the features extracted from APs by LBP-TOP and OF to select the NMPs.

Based on the advantages offered by region-based approaches, our proposed system proceeds on Regions Of Interests (ROIs) and uses deep learning techniques to extract spatiotemporal features. As can be seen in the following section, our formulation differs in that we also adopt a different label for each region when extracting the spatial features. The label vector is given based on the FACS system and the work of [51].

3. Proposed Solution

This section describes the proposed approach. An overview of our architecture is illustrated in Figure 6. This work aims to learn robust spatiotemporal features from local facial regions of MiE sequence and classify them into the corresponding labels.

3.1. Region definition

The region selection is based on two steps. First, we identify the face and then we crop the 6 ROIs identified by Ekman *et al.* [16] based on the FACS. Figure 5 shows the locations of the selected regions and the included AUs.

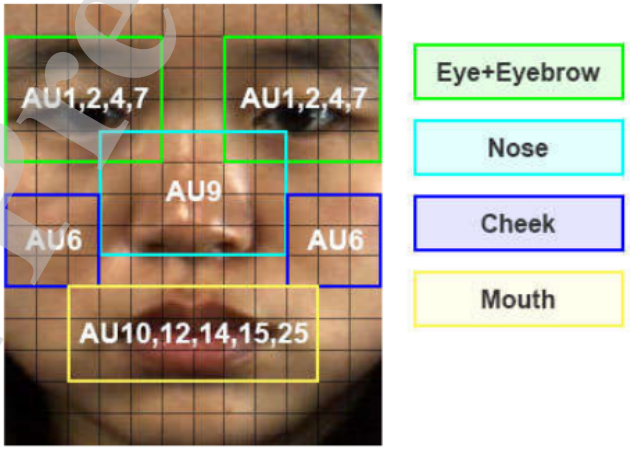

Figure 5: Illustration of the ROIs with corresponding AU [51].

To identify the face, the dlib algorithm[2] is used to detect the 68 facial landmarks. Based on these points we crop the face from the entire image. Then, two steps are needed to extract the regions from the face. First, remove all global movements from all frames by referring to the first frame of each sequence: in the used dataset the subjects are static (no global movements), thus it's unnecessary to align the faces. Secondly, we identify 6 blocks corresponding to the selected regions and crop them. Instead of one sequence S of MiE we have 6 sequences: one sequence for each region. And thus, $S = [S_r, r = 1, \ldots, R]$, where S_r is the sequence of region r and R is the number of regions (R is set to 6 in this study).

Since the used datasets have different face sizes and knowing that the size of each region is proportional to the size of the face, the size of each region will be different from

[2] http://dlib.net/face_landmark_detection.py.html

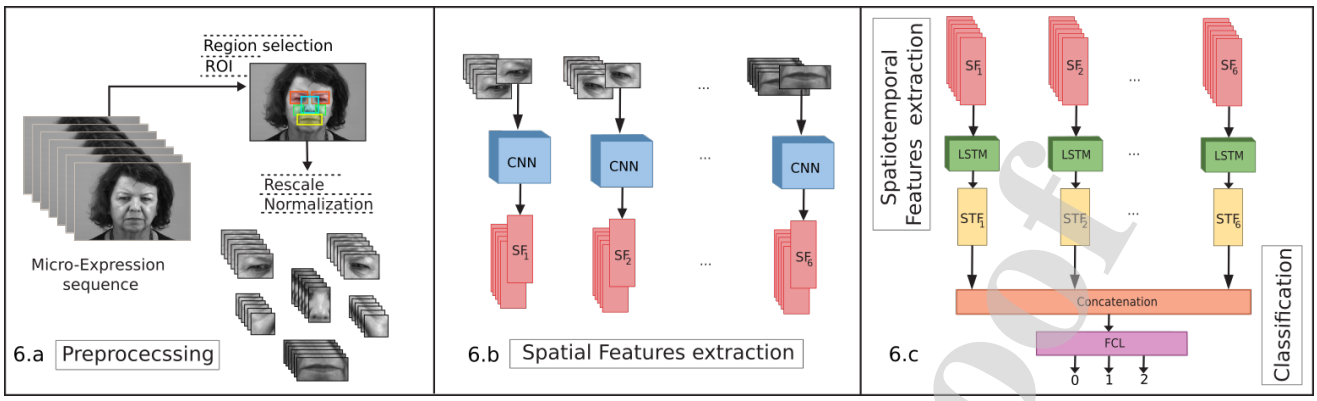

Figure 6: An overview of our proposed solution.

one dataset to another. To overcome this issue, we resize each region into a fixed size of $\{(80\times100), (80\times120), (60\times60), (60\times160)\}$, for respectively the left and the right {eye + eyebrow}, the nose, the right and left cheeks, and the mouth.

3.2. Spatial features extraction

The goal of the spatial model is to extract relevant spatial features from each region. To perform efficient training, we give a different emotion label for each region. As shown in Figure 7, a different region may express different emotional states. The regions and their corresponding labels are summarized in Table 1. If the region is responsible for that emotion, based on the work proposed by Zhao *et al.* in [51], the same emotion label is given. Otherwise, another label is given to the corresponding region, which should refer to "No Reaction" or a neutral state. It is clear, from Table 1, that the {eyes + eyebrows} and the mouth are the regions responsible for most MiEs emotions. However, the cheeks and the nose are regions that can be important for some reactions as well as to differentiate between two emotions. A drawback of using different labels for each region is the complexity to tune the hyperparameters of the model to fit all the regions.

Table 1
The ROIs and their corresponding emotions.

	eyes and eyebrows*	nose	cheeks*	mouth
Happiness	✗	✗	✓	✓
Disgust	✓	✓	✗	✓
Surprise	✓	✗	✗	✓
Repression	✗	✗	✗	✓
Sadness	✓	✗	✗	✗
Contempt	✗	✗	✗	✓

* Right and left side

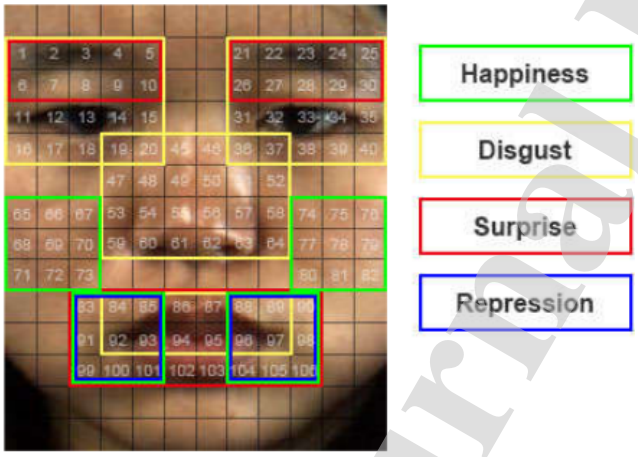

Figure 7: Identification of emotional state location [51].

Since the MiE datasets used in our experiments are small, using the proposed labels may help the model to learn faster but also the size of the network is important for this issue. According to [33, 46], deep CNN architecture may eliminate the features that are helpful for MiE recognition. Hence, the considered CNN architecture, shown in Figure 8, is much shallower (only 6 convolutional layers) than state-of-the-art models relying on InceptionNet [41] or ResNet [21].

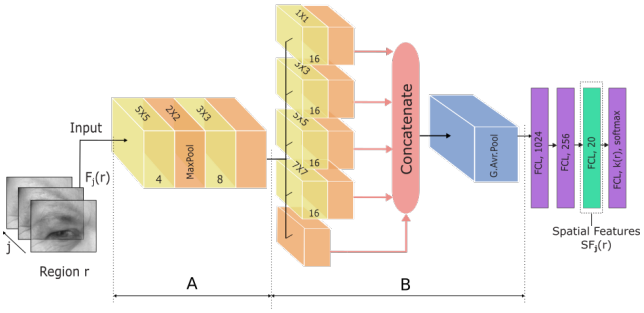

Figure 8: The architecture of the proposed shallow CNN inspired from the InceptionNet [41].

The proposed CNN model consists of two main parts. The first part *A* includes two sequential blocks, the first block is composed of 4 convolution layers with a filter size of 5×5 and the second block consists of 8 convolution layers with a filter size of 3×3. The second part *B* is composed of a max-pooling layer and 4 parallel blocks of 16 convolution layers

each, with filter size, respectively, 1×1, 3×3, 5×5 and 7×7. Each convolution block is followed by a max-Pooling layer. The outputs of the last parallel max-pooling layers are then concatenated and fed to an average-pooling layer, followed by FCLs to perform the training of the convolution part. The output of the average-pooling layer represents the spatial features. However, to reduce the size of the spatial features, we take the output of the third layer of the FCL, which contains only 20 neurons.

In our formulation, we denote the input $F_j(r)$ as the frame j of the sequence S_r, $Conv_a^b$ as the convolution operation with a filters of size $b \times b$ ($Conv_a^0 = Identity$), MP for the max-pooling layer with size 2×2, GAP for the global average pooling layer, and FCL_d for FCL with d neurons. The output Out_{S_r} of the spatial model is defined in Equation (2):

$$\begin{aligned} Out_{S_r} &= FCL_{k(r)}(SF_j(r)), \\ SF_j(r) &= FCL_{20}(FCL_{250}(FCL_{1024}(B(F_j(r))))), \\ B &= GAP(\oplus MP(Conv_{16}^i(A)), i \in [0,1,3,5,7]), \\ A &= MP(Conv_8^3(MP(Conv_4^5)), \end{aligned} \tag{2}$$

where $k(r)$ is the number of classes of the region r, $SF_j(r)$ indicates the spatial features to be saved of the j^{th} frame of the region r, and $\oplus$ is the concatenate operation.

After each convolutional or fully connected layer, we use Rectified Linear Unit (ReLU) as an activation function except for the last FCL we use Softmax. Since the model contains FCLs, it is more vulnerable to the overfitting problem. Overfitting is detected when the performance of the model on the test set is too far from the performance on the train set. To solve this problem, a Dropout layer is used. This latter chooses randomly a percentage α of neurons to be ignored during the training which forces the deep learning model to learn more robust features that are useful in conjunction with many different random subsets of the other neurons. In this work, we use Dropout with $\alpha = 0.5$.

To train the spatial model for each region as a multi-label classification problem we used focal loss (FL) [26] instead of categorical cross-entropy. The loss for the training is performed by Equation (3):

$$\mathcal{L} = FL(Softmax(FCL_{k(r)}(SF(r))), GT(r)), \tag{3}$$

where $SF(r)$ and $GT(r)$ are, respectively, the extracted spatial features for all the frames and the ground truth label of region r.

One advantage of such loss is that it deals with the problem of high unbalanced classes. The Categorical Cross-Entropy (CCE) loss function for multi-label classification is expressed in Equation (4):

$$CE(p, y) = -\sum_i y_i \log(p_i), \tag{4}$$

where y_i specifies the ground truth class and $p_i \in [0, 1]$ is the model estimated probability for the class with label y_i. The modified Cross-Entropy (CE) or the focal loss function FL is given by Equation (5):

$$FL(p, y) = -\sum_i \alpha_i y_i (1 - p_i)^\gamma \log(p_i). \tag{5}$$

In the above $\alpha_i \in [0, 1]$ is a weighting factor for class i set by inverse class frequency to contribute the imbalance between classes but it does not differentiate between hard and easy classification task of samples. That's why another modulating factor $(1 - p_i)^\gamma$ is introduced to the CE loss function, where the $\gamma >= 0$ factor is called the focusing parameter.

3.3. Temporal model

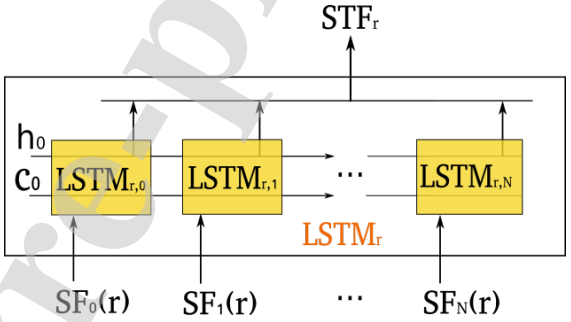

Figure 9: Temporal model: the architecture of the LSTM applied on a sequence of spatial features (SF) of the region r. c_0, h_0 : cells initialization. More details of the LSTM are in [23].

The goal of the temporal model, presented in Figure 9, is to extract spatiotemporal features $STF(r)$ (Equation (6)) from a sequence of spatial features ($SF_j(r), j \in [1..N]$) of the region r. The proposed temporal model is LSTM, and thus we have to pad with zeros all the input sequences to the same length N.

$$STF_r = LSTM_r(SF_1(r), .., SF_N(r)). \tag{6}$$

As an activation function for the output of the LSTM, a Leaky Rectified Linear Unit (LeakyReLU) is used. After that, a dropout is used with $\alpha = 0.2$ to protect the model from overfitting.

Finally, we concatenate all the extracted spatiotemporal features from all the regions and feed them to the classification network. This latter is composed of a FCL with 256 neurons, followed by LeakyReLU and a dropout with $\alpha = 0.5$, then another FCL with K neurons followed by a $softmax$ activation function. The output Out_C of the classification network is expressed by Equation (7).

$$Out_C = FCL_K(FCL_{256}(\oplus STF_r, r \in [1, \ldots, R])). \tag{7}$$

The classification and the temporal model are trained together with also the focal loss function.

Table 2
Summary of three publicly available datasets (SMIC, CASME II, SAMM) of spontaneous MiEs.

	SMIC	CASME II	SAMM
# Sequences	164	247	159
Participants	16	26	32
Resolution	640×480	640×480	2040×1088
Face	190×230	280×340	400×400
Frame rate	100	200	200
FACS coded	No	Yes	Yes
# Classes	3	5	7
Mean Age	26.7	22.03	33.24
Ethnicities	3	1	13

4. Experimental Results

In this section, we assess the performance of the proposed MiE recognition model. We start by introducing the datasets, then we ablate various design choices to show the contributions of the used model's parameters. Next, we present the results under different protocols and conditions to validate the performance of the model and finally we compare our method to the existing state-of-the-art methods.

4.1. Datasets

Experiments are conducted on three benchmark datasets of spontaneous MiE including SMIC [25], CASME II [47] and SAMM [7]. These datasets are publicly available and used on the state-of-the-art proposed solutions. Hereafter, we report details of each datasets including size, number of classes, participants and video resolution.

4.1.1. SMIC

The Spontaneous Micro-Expression (SMIC) dataset includes spontaneous micro-expressions elicited by emotional movie clips. It contains 164 sequences from 16 different subjects recorded by a high-speed (HS) camera at 100 frames per second (fps). The sequences are recorded also with a normal speed camera at 25 fps of both visual (VIS) and near-infrared (NIR) light range. So we have three datasets referred as SMIC-HS, SMIC-VIS, and SMIC-NIR for the same MiE recorded under different conditions. SMIC provides sequences with a facial resolution of (190×230) and corresponding to only 3 classes: Negative, Positive, and Surprise.

4.1.2. CASME II

The Chinese Academic of Science Micro-Expressions II (CASME II) dataset is the largest dataset of spontaneous MiE with 247[3] sequences from 35 participants with 5 classes: Happiness, Disgust, Repression, Surprise and Sadness, plus the Other class. The sequences are recorded with high temporal and spatial resolutions of 200 fps and 280×340, respectively.

[3] 247 samples was reported by the authors [47] while in the publicly available dataset, the number of samples is about 255 samples.

Table 3
Summary of the MEGC 2019 conditions to get the FULL dataset [39]

Emotion Class	SMIC	CASME II	SAMM	FULL
Negative	70	88[†]	92[⊕]	250
Positive	51	32	26	109
Surprise	43	25	15	83
Total	164	145	133	442

[†] Negative class of CASME II consists of samples from its original emotion classes of Disgust and Repression,
[⊕] Negative class of SAMM consists of samples from its original emotion classes of Anger, Contempt, Disgust, Fear and Sadness.

4.1.3. SAMM

The Spontaneous Micro-Facial Movement (SAMM) has the largest amount of different ethnicities (13 ethnicities) and age distribution (Mean age = 33.24). The video sequences are recorded with a high-resolution camera with 200 fps and contain the 7 basic classes of emotion: Disgust, Surprise, Happiness, Fear, Anger, Contempt, and Sadness. The dataset includes 159 sequences from 32 participants. It has the highest spatial resolution (400×400 pixels) among these three datasets. Further, this dataset's focus is not on the emotional labels but on the objective AUs so that all the sequences are FACS-coded and include the Onset, Apex, and Offset frames.

Table 2 summarizes the three datasets of spontaneous MiEs.

These datasets are very interesting since they afford close cases to reality with spontaneous MiEs sequences. However, they are small datasets compared to Extended Cohn-Kanade (CK+) or FER-2013 datasets for MaE recognition, that's why they are combined to form a bigger dataset.

4.2. Experimental settings

Since the three considered datasets have different conditions (number of classes, frame rate, dimensions), to combine them into a larger dataset, we used two proposed methods: with 3 classes based on the MEGC 2019 [39] test conditions and with 5 classes based on AUs [8]. Table 3 summarizes the MEGC conditions, where the number of classes is reduced to 3 (negative, positive, and surprise) to create a new larger dataset, named *FULL*, that combines SMIC, SAMM and, CASME II. This dataset has 442 sequences with 250 for negative class, 109 for positive class and 83 for surprise. The conditions for the distribution of the 5 classes based on AUs are described in [8] and the obtained distribution of classes is summarized in Table 4. The new distribution of classes is used to create a new dataset by fusing only CASMEII and SAMM. The fusion is done after reassigning the sequences of CASMEII and SAMM datasets to the 7 AU classes presented in Table 4. However, only 5 classes are preserved, the other two classes (VI and VII) are removed because the VI class contains only a very few samples and the "Others" class (VII) denotes movements that are not suited for the other cat-

Table 4
The total number of movements assigned to the new classes for both SAMM and CASME II.

Class	CASME II	SAMM	Total
I	25	24	49
II	15	13	28
III	99	20	119
IV	26	8	34
V	20	3	23
VI	1	7	8
VII	69	84	153
Total	255	159	414

Table 5
Evaluation of the model with different regions on the FULL dataset.

regions	whole face	no cheeks	no nose	all regions
Accuracy	0.85	0.89	0.91	**0.92**
UAR	0.84	0.84	0.84	**0.89**
UF1	0.84	0.84	0.85	**0.89**

egories and referred to no emotion state.

In this evaluation, 4 metrics are used including accuracy, F1-score, UAR, and UF1. The UF1 and the UAR are given by Equations (8) and (9), respectively.

$$UF1 = \frac{1}{C}\sum_{c=1}^{C} F1_c, \qquad F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \tag{8}$$

where TP_c , FP_c and FN_c are respectively true positive, false positive and false negative for class c and C is the number of classes.

$$UAR = \frac{1}{C}\sum_{c=1}^{C} ACC_c, \qquad ACC_c = \frac{TP_c}{N_c} \tag{9}$$

where N_c is the number of samples and ACC_c is the accuracy rate of class c.

4.3. Ablation study

Hereafter is presented an analysis to demonstrate the contribution of each component in the global performance of the proposed method. The experiments of this study are conducted using Leave-One-Subject-Out Cross-Validation (LOSO-CV) evaluation method under the MEGC conditions.

4.3.1. Effects of the ROIs selection

To demonstrate the importance of the selected regions we test the model with the whole face, with only some regions, and with all the regions. Table 5 gives the result of the test with different regions and proves the importance of the selected areas. It shows that using all the regions outperforms the other configurations with the whole face, without cheeks, or without the nose. We have gained about 7% accuracy using regions compared to the whole face and that's due to the two following facts: the focus on the relevant areas for MiE recognition and the use of specific labels for each selected region. From Table 5, we can also conclude that the less are the regions that contain an emotion the more those regions are important. For example, disgust emotion can be detected in the nose but also in different regions like the eyes, and the mouth, contrary to the happiness which can be expressed only with the mouth and cheeks. Therefore, the mouth and the cheeks are more important than the nose resulting in lower accuracy without cheeks than without the nose.

Table 6
Evaluation of the proposed model depending on the CNN architecture on the FULL dataset.

Metrics / CNN	Simple CNN	proposed model
Accuracy	0.85	**0.92**
UAR	0.79	**0.89**
UF1	0.79	**0.89**

Table 7
Evaluation of the proposed model depending on loss function on the FULL dataset.

Loss Function	CE	Focal Loss: $\alpha = 1/2, \gamma = 2$ [26]
Accuracy	0.91	**0.92**
UAR	0.87	**0.89**
UF1	0.87	**0.89**

4.3.2. Impact of the proposed CNN

In the proposed solution, the second contribution is the use of a special model of CNN as described in Section 3. Contrary to classical CNN models, we used a shallower architecture inspired by the architecture of the InceptionNet [41]. Table 6 compares the performance of the proposed model with a classic CNN illustrated in Figure 10. It shows that the proposed CNN model outperforms the classic one.

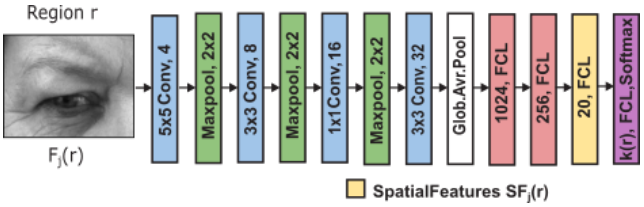

Figure 10: The architecture of a classic CNN for the left eye region.

4.4. Loss function

As a loss function for spatial and temporal models, we use two different functions: categorical cross-entropy loss and focal loss [26]. Table 7 gives the result with the categorical cross-entropy loss and with focal loss. The results with

Table 8
LOSO-CV performance of the proposed method, baselines and the recent methods (* references from the MEGC 2019 challenge)

Models (**bold**: 1st ; blue: 2nd)	FULL		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [39]*◇	0.58	0.57	0.20	0.52	0.70	0.74	0.39	0.41
Bi-WOOF [28]*◇	0.62	0.62	0.57	0.58	0.78	0.80	0.52	0.51
OFF-ApexNet [17]†	0.71	0.70	0.68	0.66	0.87	0.86	0.54	0.53
Micro-Attention [43]⊕	0.50	0.49	0.47	0.46	0.53	0.51	0.40	0.34
ATNet *(Fusion)* [34]⊕	0.63	0.61	0.55	0.54	0.79	0.77	0.49	0.48
Quang *et al.* [37]*⊕	0.65	0.65	0.58	0.58	0.70	0.70	0.58	0.59
Zhou *et al.* [52]*†	0.73	0.72	0.66	0.67	0.86	0.85	0.58	0.56
Liong *et al.* [27]*†	0.73	0.76	0.68	0.70	0.83	0.86	0.65	0.68
Liu *et al.* [29]*†	0.78	0.78	0.74	0.75	0.82	0.82	0.77	0.71
LFM-based (CNN+LSTM) [3]†	0.77	0.75	0.72	0.71	0.87	0.84	0.67	0.60
ICE-GAN [48]⊕	0.85	0.84	0.79	0.79	0.87	0.86	**0.85**	**0.82**
Our proposed method ⊕	**0.90**	**0.90**	**0.88**	**0.88**	**0.98**	**0.98**	0.78	0.81

◇ handcrafted approach, † hybrid approach, ⊕ deep learning approach.

Table 9
LOSO-CV performance of the proposed method and state-of-the-art based on the 5 classes of AUs on CASME II and SAMM datasets (* references from [8]).

Models (**bold**: 1st ; blue: 2nd)	CASME II & SAMM		CASME II		SAMM	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
LBP-TOP*◇	0.40	0.52	0.51	0.67	0.38	0.44
HOOF*◇	0.40	0.52	0.56	0.69	0.33	0.42
HOG3D*◇	0.27	0.43	0.51	0.69	0.22	0.34
ELRCN [19]†	0.41	0.57	-	-	-	-
Micro-Attention [43]⊕	0.66	0.76	-	-	-	-
Our proposed method ⊕	**0.83**	**0.86**	**0.84**	**0.88**	**0.82**	**0.79**

◇ handcrafted approach, † hybrid approach, ⊕ deep learning approach.

Table 10
Performance of the proposed method under the MEGC with LOSO-CV.

Dataset	UF1	UAR
FULL	0.90	0.90
SMIC part	0.88	0.88
CASME II part	**0.98**	**0.98**
SAMM part	0.78	0.81

Table 11
Performance of the proposed method in 5 classes based on AUs with LOSO-CV.

Dataset	UF1	Accuracy
CASME II + SAMM	0.83	0.86
CASME II part	**0.84**	**0.88**
SAMM part	0.82	0.79

the focal loss are better since it is more compatible with a scenario of imbalanced data among classes.

4.5. Comparison and discussions

The proposed model is evaluated with two cross-validation methods: the LOSO-CV method and the Cross dataset (CDB) method. In LOSO-CV, each folder contains more than one sequence of MiE of a person that didn't appear in the other folders. The LOSO-CV ensures subject-independent evaluation. For the cross dataset method, the training set will be from a dataset and the test set from another dataset. For example, the training maybe on CASME II and the test on SAMM. The second method is good to evaluate the model but the provided datasets are too small for the training especially when using a complex architecture of CNN and LSTM which may lead to a bad performance. Tables 10 and 11 show the performance of the proposed model with the LOSO-CV method and under the two conditions (with 3 classes based on the MEGC 2019 conditions or with 5 classes on the AU objective classes) while Tables 12 and 13 present the performance with cross dataset method.

From Table 10 and Table 11 we can notice that the best performance of our model is reached on the CASME II dataset. This can be explained by the characteristics of the CASME II dataset with 200 fps (100 fps for SMIC) and RGB color mode (Grayscale for SAMM). The same observation can be made in CDB validation (Table 12 and Table 13) where the proposed method outperforms other cases when CASME II is the test set that can support the hypothesis. However, when the CASME II is used as the training set (Table 12), we can notice better performance on SAMM, which is a 200 fps and grayscale dataset, than on SMIC, which is 100 fps and

Table 12
Cross dataset performance of the proposed method under the MEGC challenge.

Train set	Test set	Accuracy	F1-score
CASME II & SMIC	SAMM	0.89	0.83
CASME II & SAMM	SMIC	0.75	0.75
SMIC & SAMM	CASME II	0.92	0.91
CASME II	SMIC	0.70	0.71
CASME II	SAMM	0.87	0.79
SMIC	CASME II	0.75	0.79
SMIC	SAMM	0.79	0.75
SAMM	CASME II	0.90	0.88
SAMM	SMIC	0.75	0.76

Table 13
Cross dataset performance of the proposed method in 5 classes based on AU.

Train set	Test set	Accuracy	F1-score
CASME II	SAMM	0.75	0.59
SAMM	CASME II	**0.82**	**0.69**

RGB dataset. This latter can be an argument to explain how the temporal information is more beneficial for MiE recognition than the color scale of the image which is a part of the spatial information.

The proposed solution has been compared with other MiE recognition algorithms in different conditions. The comparison is summarized in Tables 8 and 9 which show that the proposed framework performs the best among different approaches. Table 8 shows the performance of our method, with LOSO-CV and under MEGC conditions, comparing with baselines and recent methods, that obtain high accuracy on the four datasets. The results indicate that our method outperforms handcrafted, deep learning, and hybrid approaches for almost all datasets (ICE-GAN [48] has the highest performance on SAMM).

As expected, results in Table 9 with LOSO-CV and under 5 classes based AU conditions are less accurate. We moderately improve the accuracy for three datasets (CASME II & SAMM, CASME II, SAMM) compared to other state-of-the-art methods.

5. Conclusion

In this paper, we have presented a novel spatiotemporal deep learning solution for MiE recognition from local facial regions. We have reported details and tests that support the used techniques and parameters. The main contributions of the proposed solution are the use of some particular regions of the face instead of the whole face and the extraction of spatiotemporal features from these regions with a deep learning architecture that combines CNN and LSTM. The solution is tested on different datasets and different distributions of classes and with different evaluation methods and metrics. The experiments showed that the proposed solution has a good performance that exceeds state-of-the-art solutions.

In this paper, we focused only on the recognition task without spotting and a static method for region cropping is used. Thus, in future work, we will extend the study by developing a framework that detects the active region, spots and then recognizes the MiE.

References

[1] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets. ArXiv abs/1405.3531.

[2] Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R., 2009. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1932–1939.

[3] Choi, D.Y., Song, B.C., 2020. Facial micro-expression recognition using two-dimensional landmark feature maps. IEEE Access 8, 121549–121563.

[4] Dalal, N., Triggs, B., Schmid, C., 2006. Human detection using oriented histograms of flow and appearance, in: ECCV.

[5] Daniel, B.S., 2010. Photo : Dr. cal lightman's seven universal microexpressions .

[6] Darwin, C., Prodger, P., 1998. The expression of the emotions in man and animals. Oxford University Press, USA .

[7] Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H., 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. IEEE Transactions on Affective Computing 9, 116 – 129.

[8] Davison, A.K., Merghani, W., Yap, M.H., 2017. Objective classes for micro-facial expression recognition. arXiv preprint arXiv:1708.07549. .

[9] Devangini Patel, Hong, X., Zhao, G., 2016. Selective deep features for micro-expression recognition, in: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2258–2263.

[10] Duque, C., Alata, O., Emonet, R., Konik, H., Legrand, A., 2020. Mean oriented riesz features for micro expression classification. ArXiv abs/2005.06198.

[11] Ekman, P., 1992. Facial Expressions of Emotion: an Old Controversy and New Findings. Philosophical Transactions of the Royal Society, London , 63–69.

[12] Ekman, P., 2001. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. W.W. Norton. URL: https://books.google.fr/books?id=7I_wDDfrwCgC.

[13] Ekman, P., Friesen, W., 1978. Facial Action Coding System: Investigator's Guide. Facial Action Coding System: Investigator's Guide, Consulting Psychologists Press. URL: https://books.google.fr/books?id=7pqFtQAACAAJ.

[14] Ekman, P., Friesen, W., 2003. Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. A spectrum book, Malor Books. URL: https://books.google.fr/books?id=TukNoJDgMTUC.

[15] Ekman, P., Friesen, W.V., 1969. Nonverbal leakage and clues to deception. Psychiatry , 88 – 106.

[16] Ekman P, Friesen WV, H.J., 2002. Facial Action Coding System: The Manual on CD ROM. Salt Lake City: A Human Face.

[17] Gan, Y., Liong, S.T., Yau, W.C., Huang, Y.C., Tan, L.K., 2019. OFF-ApexNet on Micro-expression Recognition System. Signal Processing: Image Communication .

[18] Guo, C., Liang, J., Zhan, G., Liu, Z., Pietikäinen, M., Liu, L., 2019. Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition. IEEE Access 7, 174517–174530.

[19] H.-Q.Khor, See, J., Phan, R.C.W., Lin, W., 2018. Enriched long-term recurrent convolutional network for facial micro-expression recognition. 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) , 667 —- 674.

[20] Haggard, E.A., Isaacs, K.S., 1966. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. Methods of research in psychotherapy , 154 — 165.

[21] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

[22] Hjortsjo, C.H., 1969. Man's Face and Mimic Language. Studen litteratur. URL: https://books.google.fr/books?id=BakQAQAAIAAJ.

[23] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

[24] Kim, D.H., Baddar, W.J., Ro, Y.M., 2016. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. Proceedings of the 2016 ACM on Multimedia Conference (Amsterdam) , 382 — 386.

[25] Li, X., Pfister, T., Huang, X., Zhao, G., Pietika, M., 2013. A Spontaneous Micro-expression Database: Inducement, Collection and Baseline. 10th Proc Int Conf Autom Face Gesture Recognition (FG2013) Shanghai, China. DOI: 10.1109/ FG.2013.6553717 .

[26] Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P., 2018. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1doi:10.1109/TPAMI.2018.2858826.

[27] Liong, S., Gan, Y.S., See, J., Khor, H., Huang, Y., 2019. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1–5. doi:10.1109/FG.2019.8756567.

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

[29] Liu, Y., Du, H., Zheng, L., Gedeon, T., 2019. A neural micro-expression recognizer. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) .

[30] Liu, Y., Zhang, J., Yan, W., Wang, S., Zhao, G., Fu, X., 2016. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing 7, 299–310.

[31] Lu, H., Kpalma, K., Ronsin, J., 2018. Motion descriptors for micro-expression recognition. Signal Processing: Image Communication 67, 108 – 117. URL: http://www.sciencedirect.com/science/article/pii/S0923596518303540, doi:https://doi.org/10.1016/j.image.2018.05.014.

[32] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101. doi:10.1109/CVPRW.2010.5543262.

[33] Nguyen, T.N., Meunier, J., 2019. Anomaly detection in video sequence with appearance-motion correspondence, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1273–1283. doi:10.1109/ICCV.2019.00136.

[34] Peng, M., Wang, C., Bi, T., Shi, Y., Zhou, X., Chen, T., 2019. A novel apex-time network for cross-dataset micro-expression recognition, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–6. doi:10.1109/ACII.2019.8925525.

[35] Pfister, T., Li, X., Zhao, G., Pietikäinen, M., 2011. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 868–875. doi:10.1109/ICCVW.2011.6130343.

[36] Polikovsky, S., Kameda, Y., Ohta, Y., 2009. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor, in: ICDP.

[37] Quang, N.V., Chun, J., Tokuyama, T., 2019. CapsuleNet for Micro-Expression Recognition. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) .

[38] Reddy, S., Karri, S.T., Dubey, S., Mukherjee, S., 2019. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. 2019 International Joint Conference on Neural Networks (IJCNN) , 1–8.

[39] See, J., Yap, M.H., Li, J., Hong, X., Wang, S.J., 2019. MEGC 2019 – The Second Facial Micro-Expressions Grand Challenge. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) .

[40] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

[41] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

[42] Verma, M., Vipparthi, S.K., Singh, G., Murala, S., 2020. Learnet: Dynamic imaging network for micro expression recognition. IEEE Transactions on Image Processing 29, 1618–1627. doi:10.1109/TIP.2019.2912358.

[43] Wang, C., Peng, M., Bi, T., Chen, T., 2020. Micro-attention for micro-expression recognition. Neurocomputing 410, 354 – 362. URL: http://www.sciencedirect.com/science/article/pii/S0925231220309711, doi:https://doi.org/10.1016/j.neucom.2020.06.005.

[44] Wang, Y., See, J., Phan, R.C.W., Oh, Y.H., 2014. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: Asian conference on computer vision, Springer. pp. 525–537.

[45] Xia, Z., Hong, X., Gao, X., Feng, X., Zhao, G., 2020. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. IEEE Transactions on Multimedia 22, 626–640. doi:10.1109/TMM.2019.2931351.

[46] Xia, Z., Peng, W., Khor, H.Q., Feng, X., Zhao, G., 2020. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Transactions on Image Processing 29, 8590–8605.

[47] Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X., 2014. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. PloS one 9.

[48] Yu, J., Zhang, C., Song, Y., Cai, W., 2020. Ice-gan: Identity-aware and capsule-enhanced gan for micro-expression recognition and synthesis. ArXiv abs/2005.04370.

[49] Zhao, G., Pietikainen, M., 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis & Machine Intelligence , 915–928.

[50] Zhao, Y., Xu, J., 2018. Necessary morphological patches extraction for automatic micro-expression recognition. Applied Sciences 8, 1811.

[51] Zhao, Y., Xu, J., 2019. An improved micro-expression recognition method based on necessary morphological patches. Symmetry 11, 497.

[52] Zhou, L., Mao, Q., Xue, L., 2019. Dual-inception network for cross-database micro-expression recognition. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) .

Author Statement

Mouath Aouayeb	Software, Methodology and Conceptualization and Writing - Original Draft
Wassim Hamidouche	Conceptualization, Methodology and Writing - Review & Editing
Catherine Soladie	Supervision and Writing - Review & Editing
Kidiyo Kpalma	Supervision
Renaud Seguier	Supervision