



HAL
open science

Cost-Effective CNNs for Real-Time Micro-Expression Recognition

Reda Belaiche, Yu Liu, Cyrille Migniot, Dominique Ginhac, Fan Yang

► **To cite this version:**

Reda Belaiche, Yu Liu, Cyrille Migniot, Dominique Ginhac, Fan Yang. Cost-Effective CNNs for Real-Time Micro-Expression Recognition. Applied Sciences, 2020, 10 (14), pp.4959. 10.3390/app10144959 . hal-02940372

HAL Id: hal-02940372

<https://hal.science/hal-02940372>

Submitted on 26 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

Cost-effective CNNs for real-time Micro-Expression recognition

Reda Belaiche*, Yu Liu, Cyrille Migniot, Dominique Gin hac and Fan Yang

ImViA EA 7535, Univ. Bourgogne Franche-Comté, Dijon, France

* Correspondence: Reda.Belaiche@u-bourgogne.fr

Version June 5, 2020 submitted to Appl. Sci.

Abstract: Micro-Expression (ME) recognition is a hot topic in computer vision as it presents a gateway to capture and understand human's daily emotions. It is nonetheless a challenging problem due to the fact ME typically being transient (lasting less than 200 ms) and subtle. Recent advances in machine learning enable new and effective methods to be adopted for solving diverse computer vision tasks. In particular, the use of deep learning techniques on large datasets outperforms classical approaches based on classical machine learning which rely on hand-crafted features. Even though available datasets for spontaneous ME are scarce and much smaller, using off-the-shelf Convolutional Neural Networks (CNNs) still demonstrates satisfactory classification results. However, these networks are heavy in terms of memory consumption and computational resources. This poses great challenges when deploying CNN-based solutions in many applications such as driver's monitoring or comprehension recognition in virtual classrooms, which demand fast and accurate recognition. As these networks are initially designed for tasks of different domains, they are over-parameterized and need to be optimized for ME recognition.

In this paper, we propose a new network based on the well-known ResNet18 which we optimize for ME classification in two ways. Firstly, we reduce the depth of the network by removing residual layers. Secondly, we introduce a more compact representation of optical flow used as input to the network. We present extensive experiments and demonstrate that the proposed network obtains accuracies comparable to the state-of-the-art methods while significantly reducing the necessary memory space. Our best classification accuracy reaches 60.17% on the challenging composite dataset containing 5 objectives classes. Our method takes only 24.6 ms for classifying a ME video clip (less than the occurrence time of the shortest ME which lasts 40 ms). Our CNN design is suitable for real-time embedded applications with limited memory and computing resources.

Keywords: computer vision, deep learning, optical flow, micro facial expressions, real-time processing.

1. Introduction

Emotion recognition has received much attention in the research community in recent years. Among the several sub-fields of emotion analysis, studies of facial expression recognition are particularly active [1–3]. In contrast to the traditional macro-expression, people are less familiar with micro facial expressions [4,5], and even fewer know how to capture and recognize them. Micro-Expression (ME) is a rapid and involuntary facial expression that exposes a person's true emotion [6]. These subtle expressions usually take place when a person conceals his or her emotions in one of the two scenarios: conscious suppression or unconscious repression. Conscious suppression happens when an individual deliberately prevents oneself from expressing genuine emotions. In contrary, unconscious repression occurs when the subject is not aware of his or her true emotions. In both cases, MEs reveal the subject's true emotions regardless of the subject's awareness. Intuitively,

ME recognition has a vast number of potential applications across different sectors, such as security field, neuromarketing [7], automobile drivers' monitoring [8] and lies and deceit detection [5].

Psychological research shows that facial MEs generally are transient (e.g., remaining less than 200 ms) and very subtle [9]. The short duration and subtlety incur great challenges for human to perceive and recognize them. To enable better ME recognition by human, Ekman and his team developed the ME training tool (METT). Even with the help of this training tool, human can barely achieve around 40% accuracy [10]. Moreover, human's decisions are prone to be influenced by individual's perception varied along different subjects and time, resulting in less objective results. Therefore, a bias-free and high-quality automatic system for facial ME recognition is highly sought after.

A number of earlier solutions to automate facial ME recognition has been based on geometry or appearance feature extraction methods. Specifically, geometric-based features encode geometric information of the face, such as shapes and locations of facial landmarks. On the other hand, appearance-based features describe the skin texture of faces. Most existing methods [11,12] attempt to extract low-level features such as the widely used Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) [13–15] from different facial regions, and simply concatenate them for ME recognition. Nevertheless, transient and subtle ME inherently makes it challenging for low level-features to effectively capture essential movements in ME. At the same time, these features can also be affected by irrelevant information or noise in video clips, which further weakens their discrimination capabilities especially on inactive facial regions with less dynamics [16].

Recently, more approaches based on mid-level and high-level features have been proposed. Among these methods, the pipeline composed of optical flow and deep learning has demonstrated its high effectiveness for MEs recognition in comparison with traditional ones. The studies applying deep learning to tackle the ME classification problem usually considered well-known Convolutional Neural Networks (CNNs) such as ResNet [17] and VGG [18]. These studies re-purpose the use of off-the-shelf CNNs by giving them input data token from the optical flow extracted from the MEs. While achieving good performance, these neural networks are quite heavy in terms of memory usage and computation.

In specific applications, for example during automobile drivers' monitoring or students' comprehension recognition in virtual education systems, fast and effective processing methods are necessary to capture emotional responses as quickly as possible. Meanwhile, thanks to great progresses in parallel computing, parallelized image processing devices such as embedded systems are easily accessible and affordable. Already well-adopted in diverse domains, these devices possess multiple strengths in terms of speed, embeddability, power consumption and flexibility. These advantages however are often at the cost of limited memory and computing power.

The objective of this work is to design an efficient and accurate ME recognition pipeline for embedded vision purpose. First of all, our design takes into account thorough investigations on different CNN architectures. Next, different optical flow representations for CNN inputs have been studied. Finally, our proposed pipeline achieves competitive accuracy for ME recognition as state-of-the-art approaches while being real-time capable and using less memory. The paper is organized as follows. In Section 2, several recent related work are reviewed. Section 3 explains the proposed methodology in order to establish cost-effective CNNs for fast ME recognition. Section 4 provides experimental results and performance evaluations. Lastly, Section 5 concludes the paper.

2. Related works

MEs begin at the onset (first frame where the muscles of the facial expressions start to contract), finish at the offset (last frame, where the face returns to its neutral state), and reach their pinnacle at the apex frames (see Figure 1). Because of their very short duration and low intensity, ME recognition and analysis are considered as difficult tasks. Earlier studies proposed using low-level features such as LBP-TOP to address these problems. LBP-TOP is a 3D descriptor extended from the traditional 2D LBP. It encodes the binary patterns between image pixels, and the temporal relationship between pixels and their neighboring frames. The resulting histograms are then concatenated to represent the

85 temporal changes over entire videos. LBP-TOP has been widely adopted in several studies. Pfister et
 86 al. [13] applied LBP-TOP for spontaneous ME recognition. Yan et al. [14] achieved 63% ME recognition
 87 accuracy on their CASME II database using LBP-TOP. In addition, LBP-TOP has also been used to
 88 investigate differences between micro-facial movement sequences and neutral face sequences.

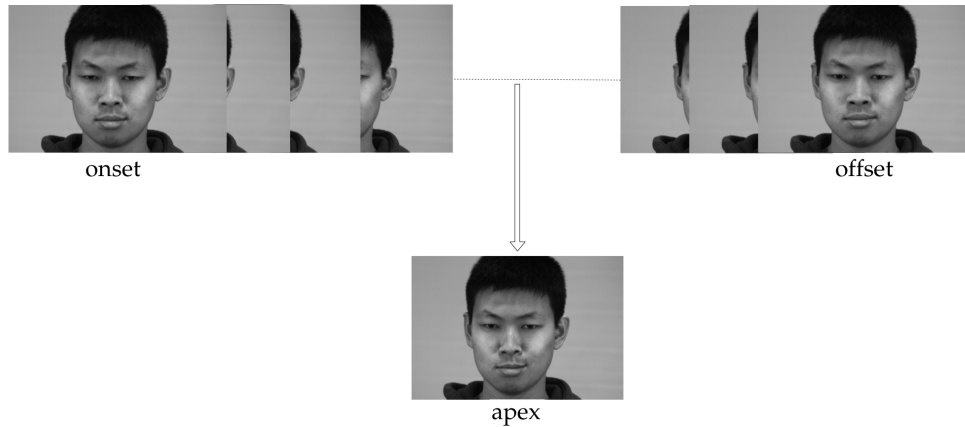


Figure 1. Example of a ME: the maximum movement intensity occurs at the apex frame.

89 Several studies aimed to extend low-level features extracted by LBP-TOP as they still could not
 90 reach satisfactory accuracy. For example, Liong et al. [19] proposed to assign different weights to
 91 local features, putting more attention on active facial regions. Wang et al. [11] studied the correlation
 92 between color and emotions by extracting LBP-TOP from the tensor independent color space (TICS).
 93 Ruiz-Hernandez and Pietikäinen [20] used the re-parameterization of second order Gaussian jet on
 94 the LBP-TOP, achieving promising ME recognition result on the SMIC database [21]. Considering
 95 that LBP-TOP consists of redundant information, Wang et al. [22] proposed the LBP-Six Intersection
 96 Points (LBP-SIP) method which is computationally more efficient and achieves higher accuracy on the
 97 CASEME II database. We also note that the STCLQP (SpatioTemporal Completed Local Quantization
 98 Patterns) proposed by Huang et al. [23] achieved a substantial improvement for analyzing facial MEs.

99 Over the years as research shows that it is non-trivial for low-level features to effectively capture
 100 and encode ME's subtle dynamic patterns (especially from inactivate regions), other methods shift to
 101 exploit mid- or high-level features. He et al. [16] developed a novel multi-task mid-level feature
 102 learning method to enhance the discrimination ability of the extracted low-level features. The
 103 mid-level feature representation is generated by learning a set of class-specific feature mappings.
 104 Better recognition performance has been obtained with more available information, features with
 105 better discrimination and generalization abilities. A simple and efficient method known as Main
 106 Directional Mean Optical-flow (MDMO) was employed by Liu et al. [24]. They used optical flow to
 107 measure the subtle movement of facial regions of interest (ROIs) that were spotted based on the Facial
 108 Action Coding System (FACS). Oh et al. [25] also applied the monogenic Riesz wavelet representation
 109 in order to amplify subtle movements of MEs.

110 The aforementioned methods indicate that the majority of existing approaches heavily rely on
 111 hand-crafted features. Inherently, they are not easily transferable as the process of feature crafting and
 112 selection depend heavily on domain knowledge and researchers' experience. In addition, methods
 113 based on hand-crafted features are not accurate enough to be applied in practice. Therefore, high-level
 114 feature descriptors which better describe different MEs and can be automatically learned are desired.
 115 Recently, more and more vision-based tasks have shifted to deep CNN-based solutions due to their
 116 superior performance. Recent developments in ME recognition are also inspired by these advancements
 117 by incorporating CNN models within the ME recognition framework.

118 Peng et al. [26] proposed a two-stream convolutional network DTSCNN (Dual Temporal Scale
119 Convolutional Neural Network) to address two aspects: overfitting problem caused by small sizes of
120 existing ME databases and use of high-level features. We can observe four characteristics of DTSCNN:
121 (i) separate features were first extracted from ME clips from two shallow networks and then fused;
122 (ii) data augmentation and higher drop-out ratio were applied in each network; (iii) two databases
123 (CASME I and CASME II) were combined to train the network; (iv) the data fed to the networks were
124 optical-flow images instead of raw RGB frames.

125 Khor et al. [27] studied two variants of an Enriched LRCN (Long-term Recurrent Convolutional
126 Network) model for ME recognition. Spatial enrichment (SE) refers to channel-wise stacking of
127 gray-scale and optical flow images as new input to CNN. On the other hand, temporal enrichment
128 (TE) stacks obtained features. Their TE model achieves better accuracy on a single database, while the
129 SE model is more robust against the cross-domain protocol involving more databases.

130 Liong et al. [28] designed a Shallow Triple Stream Three-dimensional CNN (STSTNet). The
131 model takes input stacked optical flow images computed between the onset and apex frames (optical
132 strain, horizontal and vertical flow fields), followed by three shallow convolution layers in parallel
133 and a fusion layer. The proposed method is able to extract rich features from MEs while being
134 computationally light, as the fused features are compact yet discriminative.

135 Our objective is to realize a fast and high-performance ME recognition pipeline for embedded
136 vision applications under several constraints, such as embeddability, limited memory and restricted
137 computing resources. Inspired by existing works [26,28], we explore different CNN architectures and
138 several optical flow representations for CNN inputs to find cost-effective neural network architectures
139 that are capable of recognizing MEs in real-time.

140 3. Methodology

141 The studies applying deep learning to tackle the ME classification problem [29–32] usually used
142 pretrained CNNs such as ResNet [17] and VGG [18] and applied transfer learning to obtain ME
143 features. In our work, we first select off-the-shelf ResNet18 because it provides the best trade-off
144 between accuracy and speed on the challenging ImageNet classification and is recognized for these
145 performances in transfer learning. ResNet [17] explicitly lets the stacked layers fit a residual mapping.
146 Namely, the stacked non-linear layers are let to fit another mapping of $F(x) := H(x) - x$ where $H(x)$
147 is the desired underlying mapping and x the initial activations. The original mapping is recast into
148 $F(x) + x$ by feedforward neural networks with shortcut connections. ResNet18 has 20 convolutional
149 layers (CL) (17 successive CL and 3 branching ones). Residual links after each pair of successive
150 convolutional units are used and the kernel size after each residual link is doubled. As ResNet18 is
151 designed to extract features from RGB color images, it requires inputs to have 3 channels.

152 In order to accelerate processing speed in the deep learning domain, the main current trend in
153 decreasing complexity of CNN is to reduce the number of parameters. For example, Hui et al. [33]
154 proposed a very compact LiteFlowNet which is 30 times smaller in the model size and 1.36 times faster
155 in the running speed in comparison with the state-of-the-art CNNs for optical flow estimation. In
156 [34], Rieger et al. explored parameter-reduced residual networks on in-the-wild datasets, targeting
157 real-time head pose estimation. They experimented various ResNet architectures with a varying
158 number of layers to handle different image sizes (including low-resolution images). The optimized
159 ResNet achieved state-of-the-art accuracy with real-time speed.

160 Well known CNN is created for specific problems and therefore over calibrated when they are
161 used in other contexts. ResNet18 was made for end-to-end object recognition: the dataset used for
162 training had hundreds of thousands of images for each class and more than a thousand classes in total.
163 Based on the fact that: (i) ME recognition study considers in maximum 5 classes and the datasets of
164 spontaneous MEs are scarce and contain much fewer samples, and (ii) optical flows are high-level
165 features contrary to low-level color features and so require shallower network, we have reduced the
166 architecture of ResNet18 by iteratively removing residual layers. This allows us to assess the influence

167 of the depth of the network on its classification capacities in our context and therefore to estimate the
 168 relevant calibration of the network.

169 Figure 2 illustrates the reduction protocol: at each step the last residual layer with two CL is
 170 removed and the previous one is connected to the fully connected layer. Only networks with an odd
 171 number of CL are therefore proposed. As highlighted in Table 1 the decrease in the number of CL
 172 poses a significant impact on the number of learnable parameters of the network, which directly affects
 173 the forward propagation time.

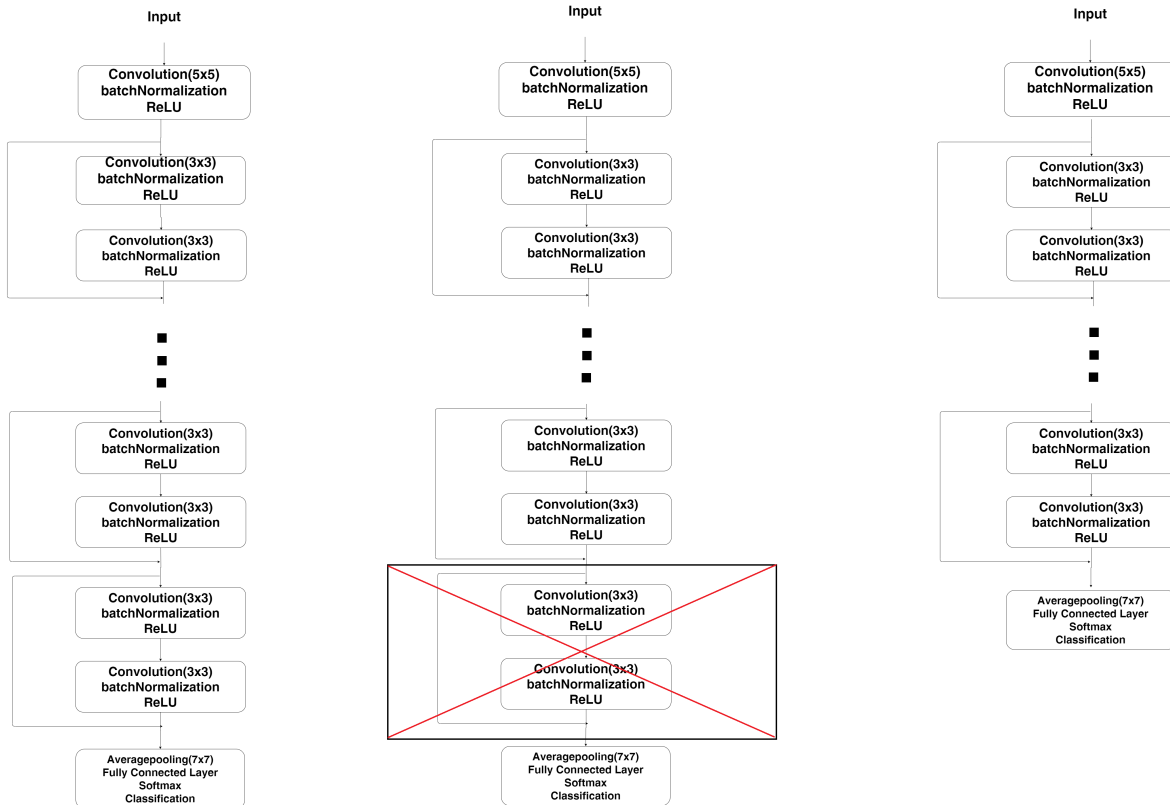


Figure 2. Depth reduction of a deep neural network: in the initial network, each residual layer contains two CL (left); the last residual layer is removed (middle) to obtain a shallower network (right).

Table 1. Number of CL and number of learnable parameters in the proposed architectures.

| CL | 17 | 15 | 13 | 11 | 9 | 7 | 5 | 3 | 1 |
|---------------|------------|-----------|-----------|-----------|---------|---------|---------|---------|--------|
| Nb. of param. | 10,670,932 | 5,400,725 | 2,790,149 | 1,608,965 | 694,277 | 398,597 | 178,309 | 104,197 | 91,525 |

174 Once the network depth has been correctly estimated, the dimension of the input has to be
 175 optimized. In our case, CNNs take optical flows extracted between the onset and apex frames of
 176 ME video clips. It is between these two moments that the motion is most likely to be the strongest.
 177 The dimensionality of inputs determines the complexity of the network that uses them since the
 178 reduction in input channels dictates the number of filters to be used throughout all following layers of
 179 the CNN. The optical flow between the onset (Figure 3-a) and the apex (Figure 3-b) typically has a
 180 3-channel representation to be used in a pretrained architecture designed for 3-channel color images.
 181 This representation however may not be optimal for ME recognition.

182 Optical flow can be described as the change of structured patterns of light between successive
 183 frames to measure movement of a pixel over a period of time. Optical flow estimation techniques are
 184 based on the assumption of brightness invariance:

$$I(x, y, t) = I(x + \delta_x, y + \delta_y, t + \delta_t) \quad (1)$$

185 where $I(x, y, t)$ is the intensity of pixel in position (x, y) at time t .

186 The optical flow is represented as a vector (Figure 3-c) indicating the direction and intensity of the
 187 motion. The projection of the vector on the horizontal axis corresponds to the V_x field (Figure 3-d)
 188 while its projection on the vertical axis is the V_y field (Figure 3-e). The magnitude M is the norm of the
 189 vector (Figure 3-f). Figure 4 illustrates this representation of one optical flow vector. The horizontal and
 190 vertical components V_x and V_y of the optical flow correspond to the spatial variation (δ_x, δ_y) obtained
 191 by minimizing the difference between the left and right term of Equation 1.

192 In this paper, the optical flow is estimated by the Horn-Schunck method [35]. This method
 193 assumes that the optical flow is smooth over the entire image. Hence minimizing the following
 194 equation estimates the velocity field:

$$E = \iint \left(\frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} \right)^2 dx dy + \alpha \iint \left(\|\nabla V_x\|^2 + \|\nabla V_y\|^2 \right) dx dy \quad (2)$$

195 α is a regularization parameter that controls the degree of smoothness and is usually selected
 196 heuristically. This energy is iteratively minimized until convergence from the following equations :

$$V_x^{k+1} = \bar{V}_x^k - \frac{\frac{\partial I}{\partial x} \left(\frac{\partial I}{\partial x} \bar{V}_x^k + \frac{\partial I}{\partial y} \bar{V}_y^k + \frac{\partial I}{\partial t} \right)}{\alpha^2 + \frac{\partial I^2}{\partial x^2} + \frac{\partial I^2}{\partial y^2}} \quad (3)$$

$$V_y^{k+1} = \bar{V}_y^k - \frac{\frac{\partial I}{\partial y} \left(\frac{\partial I}{\partial x} \bar{V}_x^k + \frac{\partial I}{\partial y} \bar{V}_y^k + \frac{\partial I}{\partial t} \right)}{\alpha^2 + \frac{\partial I^2}{\partial x^2} + \frac{\partial I^2}{\partial y^2}} \quad (4)$$

197 where \bar{V} is the weighted average of V in a neighbourhood.

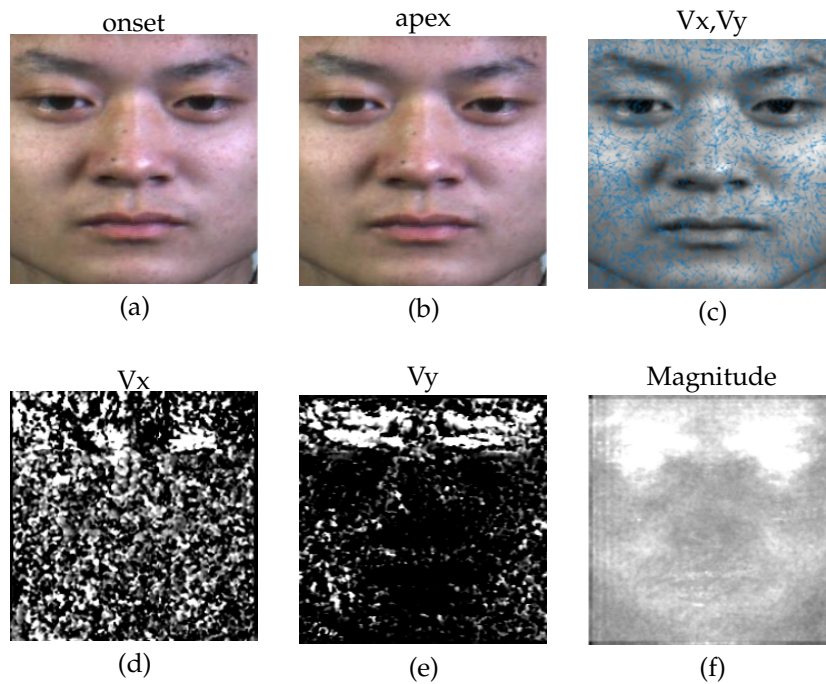


Figure 3. Optical flow is computed between the onset (a) and the apex (b): vectors obtained for a random sample of pixels (c), V_x field (d), V_y field (e) and magnitude field (f).

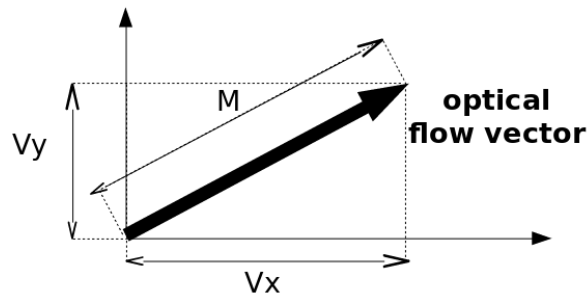


Figure 4. Visualisation of M , V_x and V_y for one optical flow vector.

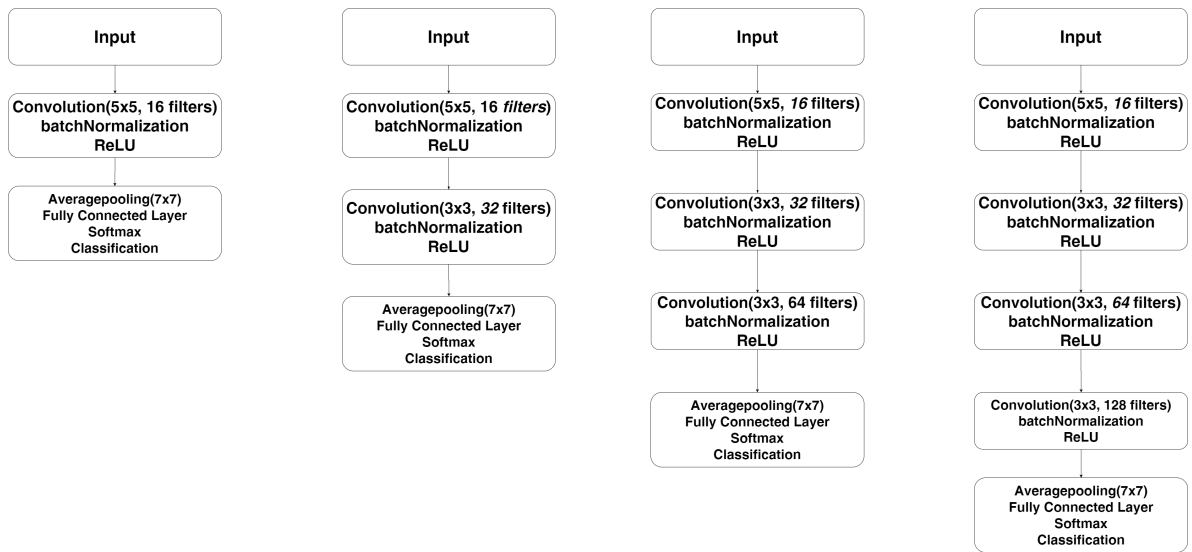


Figure 5. Proposed networks composed of one to four (from left to right) CL for various representations of the optical flow as input.

198 When classifying ME, the resulting matrices V_x , V_y and M are traditionally given as input
 199 to the CNN. Nonetheless, the third channel is inherently redundant since M is computed from V_x
 200 and V_y . Optical flow composed of the 2-channel V_x and V_y field could already provide all relevant
 201 information. Furthermore, we hypothesize that even a single channel motion field itself could be
 202 descriptive enough. Hence we have created and evaluated networks taken as input the optical flow
 203 in a two-channel representation (V_x - V_y) and in an one-channel representation (M , V_x or V_y). For
 204 this purpose, the proposed networks begin by a number of CL related to the depth optimization
 205 followed by a batch normalization and ReLU. Then the networks end by a maxpooling layer and a
 206 fully connected layer. The Figure 5 presents the architectures used with one to four CL according
 207 to the results of the experiments in Section 4. As illustrated in Table 2, a low dimensional input leads to
 208 a significant reduction in the number of learnable parameters and therefore in the complexity of the
 209 system.

Table 2. Number of learnable parameters according to the dimensionality of the input of the network.

| Input | 1 CL | 2 CL | 3 CL | 4 CL |
|----------------|---------|---------|---------|-----------|
| Single channel | 82,373 | 168,997 | 333,121 | 712,933 |
| Double channel | 165,541 | 348,005 | 709,477 | 1,620,197 |

210 4. Experiments

211 4.1. Dataset and validation protocol presentation

212 Two ME databases are used in our experiments. CASME II (Chinese Academy of Sciences
213 Micro-Expression) [14] is a comprehensive spontaneous ME database containing 247 video samples,
214 collected from 26 Asian participants with an average age of 22.03 years old. Compared to the first
215 database, the Spontaneous Actions and Micro-Movements (SAMM) [36] is a more recent one consisting
216 of 159 micro-movements (one video for each). These videos are collected spontaneously from a
217 demographically diverse group of 32 participants with a mean age of 33.24 years old and a balanced
218 gender split. Originally intended for investigating micro-facial movements, SAMM initially collected
219 the 7 basic emotions.

220 Both the CASME II and SAMM databases are recorded at a high-speed frame rate of 200 fps. They
221 also both contain "objective classes", as provided in [37]. For this reason, Facial MEs Grand Challenge
222 2018 [38] proposed to combine all samples from both databases into a single composite dataset of 253
223 videos with five emotion classes. It should be noted that the repartition is not very well balanced.
224 Namely, this composite database is composed of 19.92% "happiness", 11.62% "surprise", 47.30% "anger",
225 11.20% "disgust" and 9.96% "sadness".

226 Similar to [38], we applied the Leave One Subject Out (LOSO) cross-validation protocol for ME
227 classification, where one subject's data is used as a test set in each fold of the cross-validation. This
228 is done to better reproduce realistic scenarios where the encountered subjects are not present during
229 training of the model. In all experiments, recognition performance is measured by accuracy, which is
230 the percentage of correctly classified video samples out of the total number of samples in the database.

231 The Horn-Schunck method [35] was selected to compute optical flow. This algorithm is widely
232 used for optical flow estimation in many recent studies in virtue of its robustness and efficiency.
233 Throughout all experiments, we train the CNN models with a mini-batch size of 64 for 150 epochs
234 using the RMSprop optimization. Feature extraction and classification are both handled by the CNN.
235 Simple data augmentation is applied to double the training size. Specifically, for each ME video clip
236 used for training, in addition to the optical flow between the onset and apex frame, we also include a
237 second flow computed between the onset and apex+1 frame.

238 4.2. ResNet depth study

239 In order to find the ResNet depth which permits an optimal compromise between the ME
240 recognition performance and the number of learnable parameters, we tested different CNN depths
241 using the method described in Section 3. The obtained accuracies are given in Table 3:

Table 3. Accuracies varied by the number of convolution layers (CL) and associated number of learnable parameters.

| Nb. of CL | 17 | 15 | 13 | 11 | 9 | 7 | 5 | 3 | 1 |
|---------------|------------|-----------|-----------|-----------|---------|---------|---------|---------|--------|
| Nb. of param. | 10,670,932 | 5,400,725 | 2,790,149 | 1,608,965 | 694,277 | 398,597 | 178,309 | 104,197 | 91,525 |
| Accuracy | 57.26% | 57.26% | 60.58% | 59.34% | 60.17% | 61.00% | 58.51% | 60.17% | 58.92% |

242 We observe that the best score is achieved by ResNet8 which has seven CL. However, the scores
243 achieved by different numbers of CL do not vary much. Furthermore, beyond seven CL, adding more
244 CL doesn't improve the accuracy of the model. The fact that accuracy doesn't increase along with
245 depth confirms that multiple successive CL are not necessary to achieve a respectable accuracy. The
246 most interesting observation is that with a single CL, we achieve a score that is not very far from the
247 optimal score while the size of the model is much more concise. This suggests that instead of deep

248 learning, a more "classical" approach exploiting shallow neural networks presents an interesting field
 249 to explore when considering portability and computation efficiency for embedded systems. That is the
 250 principal reason we will restrict our study to shallow CNNs.

251 4.3. CNN input study

252 In this subsection, we study impacts of optical flow representations on ME recognition
 253 performance. Two types of CNN have been investigated, one with 1-channel input (V_x , V_y , or
 254 M) and the other one using the 2-channel V_x - V_y pair. Due to the fact that off-the-shelf CNNs typically
 255 take 3-channel inputs and are pre-trained accordingly, applying transfer learning to adapt to our
 256 models is a nontrivial task. Instead, we created custom CNNs and trained them from scratch. Table 4
 257 shows recognition accuracies of different configurations using a small number of CNN layers.

258 We can observe that the V_x - V_y pair and V_y alone give the best results, both representations
 259 achieving 60.17% accuracy. On the other hand, using magnitude alone leads to similar accuracy as
 260 those of V_y and V_x - V_y pair with a score of 59.34%. V_x gets the worst results overall, with a maximum
 261 score of 54.34%. This observation indicates that the most prominent features for ME classification
 262 might indeed be more dominant in vertical movement rather than the horizontal one. This assumption
 263 is logical when thinking about the muscle movements happening in each known facial expression.

Table 4. Accuracies under various CNN architectures and optical flow representations.

| | 1 CL | 2 CL | 3 CL | 4 CL |
|---------------|--------|--------|---------------|--------|
| V_x | 52.24% | 54.34% | 53.92% | 53.50% |
| V_y | 58.09% | 59.34% | 60.17% | 60.17% |
| V_x - V_y | 58.51% | 59.75% | 60.17% | 58.09% |
| M | 58.09% | 58.92% | 59.34% | 59.34% |

264 To better visualize the difference in the high-level features present in V_x , V_y and the Magnitude,
 265 we did an averaging on all the different samples according to their classes. The result can be seen in
 266 Figure 6. We observe that V_x exhibits a non-negligible quantity of noise. Magnitude and V_y on the
 267 other hand have clear regions of activity for each class. The regions of activity are aligned with the
 268 muscles responsible of each facial expression.

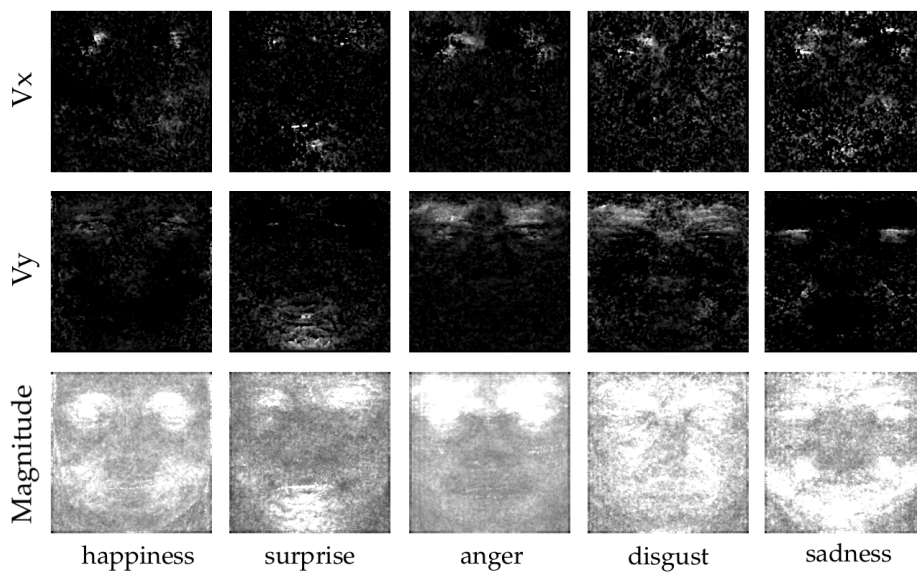


Figure 6. Average optical flow obtained in the dataset per ME class. Studied classes are in order from left to right: happiness, surprise, anger, disgust and sadness.

269 4.4. Classification analysis

270 In order to understand obtained results, we measured cosine similarity of features extracted by
 271 three CNNs: ResNet8 (Section 4.2), Vx-Vy-3 CL and Vy-3 CL (Section 4.3). Usually, the convolutional
 272 layers of CNNs are considered as different feature extractors; only the last fully connected layer directly
 273 performs the classification task. The features just before classification can be represented in vector
 274 format. Cosine similarity measures the similarity between two vectors a and b using Equation 5:

$$\text{cosine}(a, b) = \frac{a^T b}{\|a\| \|b\|} \quad (5)$$

275 Cosine similarity values fall within the range of $[-1, 1]$; values closer to 1 indicate higher similarity
 276 between two vectors. Tables 5, 6 and 7 display the cosine similarity values: with 2 samples \times 5
 277 ME classes, we calculated intra-similarity and average inter-similarity of each class using the same
 278 configuration for three CNNs.

| | happiness | surprise | anger | disgust | sadness |
|-----------|-----------|----------|--------|---------|---------|
| happiness | 0.6007 | 0.1320 | 0.0574 | 0.0146 | 0.1154 |
| surprise | 0.1320 | 0.5572 | 0.0485 | 0.0667 | 0.1415 |
| anger | 0.0574 | 0.0485 | 0.5260 | 0.0318 | 0.0698 |
| disgust | 0.0146 | 0.0667 | 0.0318 | 0.5663 | 0.0159 |
| sadness | 0.1154 | 0.1415 | 0.0698 | 0.0159 | 0.5099 |

Table 5. Cosine similarity for the 3 CL CNN with single-channel input Vy

| | happiness | surprise | anger | disgust | sadness |
|-----------|-----------|----------|--------|---------|---------|
| happiness | 0.5615 | 0.1700 | 0.1171 | 0.1155 | 0.1195 |
| surprise | 0.1700 | 0.5831 | 0.1432 | 0.1502 | 0.1618 |
| anger | 0.1171 | 0.1432 | 0.5672 | 0.1176 | 0.1503 |
| disgust | 0.1155 | 0.1502 | 0.1176 | 0.5447 | 0.1225 |
| sadness | 0.1195 | 0.1618 | 0.1503 | 0.1225 | 0.5443 |

Table 6. Cosine similarity for the 3 CL CNN with double-channel inputs (Vx-Vy)

| | happiness | surprise | anger | disgust | sadness |
|-----------|-----------|----------|--------|---------|---------|
| happiness | 0.8464 | 0.3966 | 0.3860 | 0.3126 | 0.2960 |
| surprise | 0.3966 | 0.8159 | 0.4040 | 0.3362 | 0.3324 |
| anger | 0.3860 | 0.4040 | 0.8344 | 0.3654 | 0.3307 |
| disgust | 0.3126 | 0.3362 | 0.3654 | 0.8598 | 0.2363 |
| sadness | 0.2960 | 0.3324 | 0.3307 | 0.2363 | 0.9343 |

Table 7. Cosine similarity for ResNet8

279 Firstly, we observe that diagonal values (intra-class) across all three CNNs are significantly higher
 280 in comparison with other values (inter-class). This illustrates that all three CNNs are capable to separate
 281 different ME classes. Secondly, the intra-class cosine similarity of ResNet is closer to 1, suggesting
 282 that ResNet features are more discriminative. We hypothesize that our simplified CNNs with reduced
 283 layers extract less refined features, resulting in the minor decrease in performance (61.00% vs. 60.17%).

284 4.5. Performance evaluations

285 In this subsection, we measure our proposed method on three aspects: recognition accuracy,
 286 needed memory space and processing speed. Since we obtain optimal results by using the Vy field and
 287 3-layer CNN, further evaluations will concentrate on this particular configuration.

288 **Evaluation on recognition accuracy:** we performed an accuracy comparison of 5 objective ME
 289 class recognition (see Table 8). Our best CNN reaches similar performance as those of other studies
 290 using the same protocol of validation. It is worth mentioning that Peng et al. [39] employed
 291 a macro-to-micro transferred ResNet10 model and obtained a better result. Their work used 4
 292 macro-expressions datasets (> 10K images) and some preprocessing such as colour shift, rotation
 293 and smoothing. These additional operations make their proposed method difficult for deployment on
 294 embedded systems. Seeing the confusion matrix of our model (Figure 7), we can also notice that the
 295 distribution of correct assessments for Vy is more balanced than the ones gotten from [27] (Figure 8).

296 The DTSCNN proposed by Peng et al. in [26] opted for two optical flows computed differently
 297 from a ME sample, which make the whole network robust to different frame rate of ME videos. In
 298 detail, the first optical flow is calculated using 64 frames around the apex to adapt to the frame rate of
 299 CASME I. Similarly, the second optical flow is given by the 128 frames around the apex adapted to
 300 the frame rate of CASME II. In case the number of frames composing the ME is not sufficient, a linear
 301 interpolation method is used to normalize the video clips. Their study uses two CNNs in parallel to
 302 extract two separate features before concatenating them. The resulted feature vector is then fed as input
 303 to an SVM to be classified. The DTSCNN was tested on four classes (positive, negative, surprise, and
 304 other) from a composite dataset consisting of the CASME I and CASME II databases, and it achieves
 305 an average recognition rate of 66.67%. The STSTNet proposed by Liong et al. in [28] makes use of
 306 three dimensional CNNs which carry out three dimensional convolutions instead of two-dimensional
 307 ones (such as ResNet, VGG, the networks presented in [26], [27], [39] and our study). It was tested on
 308 three classes: positive, negative, and surprise from a composite database consisting of samples from
 309 the SMIC, CASME II and SAMM databases. It achieved an unweighted average recall rate of 76.05%
 310 and an unweighted F1-score of 73.53%. Both of these two frameworks are not very suitable for real
 311 time embedded applications constrained by limited memory and computing resources.

| | happiness | surprise | anger | disgust | sadness |
|-----------|-----------|----------|-------|---------|---------|
| happiness | 43.8% | 10.4% | 20.8% | 8.3% | 16.7% |
| surprise | 10.7% | 32.1% | 35.7% | 7.1% | 14.3% |
| anger | 2.6% | 4.4% | 83.3% | 7.0% | 2.6% |
| disgust | 7.4% | 7.7% | 40.7% | 40.7% | 3.7% |
| sadness | 20.8% | 12.5% | 29.2% | 0% | 37.5% |

Figure 7. Confusion matrix corresponding to our network with 3 CL and Vy as input.

Table 8. Comparison between our method and those of other top-performers from literature.

| Method | Accuracy |
|---------------------|----------|
| <i>LBP_TOP</i> [27] | 42.29% |
| Khor et al [27] | 57.00% |
| Peng et al [39] | 74.70% |
| Proposed method | 60.17% |

| | happiness | surprise | anger | disgust | sadness |
|-----------|-----------|----------|-------|---------|---------|
| happiness | 43% | 6% | 35% | 6% | 10% |
| surprise | 21% | 29% | 43% | 0% | 7% |
| anger | 3% | 3% | 91% | 3% | 0% |
| disgust | 21% | 3% | 65% | 6% | 6% |
| sadness | 22% | 13% | 35% | 4% | 26% |

Figure 8. Confusion matrix obtained by the work of [27].

312 **Evaluation on memory space:** Table 9 summarizes the number of learnable parameters and used
 313 filters according to the dimensionality of the network inputs. The minimum required memory space
 314 corresponds to 333,121 parameter storage, which is less than 3.12% of that of off-the-shelf ResNet18.

Table 9. Number of learnable parameters and filters (in brackets) of various network architectures under different input dimensions.

| Input | 1 CL | 2 CL | 3 CL | 4 CL |
|----------------|-----------------|-----------------|------------------|--------------------|
| Single channel | 82,373 (16) | 168,997 (48) | 333,121 (112) | 712,933 (240) |
| Double channel | 165,541 (32) | 348,005 (96) | 709,477 (224) | 1,620,197 (480) |

315 **Evaluation on processing speed:** we used a mid-range computer with an Intel Xeon processor
 316 and an Nvidia GTX 1060 graphic card to carry out all the experiments. The complete pipeline is
 317 implemented in MatLAB 2018a with its deep learning toolbox. Our model which achieves the best
 318 score is the CNN with a single-channel input and three successive CL. It needs 12.8 ms to classify the
 319 vertical component V_y . The optical flow between two frames requires 11.8 ms to compute using our
 320 computer, leading to a total runtime to classify an ME video clip of 24.6 ms. In our knowledge, the
 321 proposed method outperforms most ME recognition systems in terms of processing speed.

322 5. Conclusion and future works

323 In this paper, we propose cost-efficient CNN architectures to recognize spontaneous MEs. We
 324 first investigated the depth of the well-known ResNet18 network to demonstrate that using only a
 325 small number of layers is sufficient in our task. Based on this observation, we have experienced several
 326 representations at network's input.

327 Following several previous studies, we fed CNNs with optical flow estimated from the onset and
 328 apex of MEs. Different flow representations (horizontal V_x , vertical V_y , Magnitude M and V_x - V_y pair)
 329 have been tested and evaluated on a composite dataset (CASME II and SAMM) for recognition of five
 330 objective classes. The results obtained on the V_y input alone are more convincing. It is likely due to the
 331 fact that such an orientation is more suitable describing ME's motion and its variations between the
 332 different expression classes. Experimental results demonstrated that the proposed method can achieve
 333 similar recognition rate when compared with state-of-the-art approaches.

334 Finally, we obtained an accuracy of 60.17% with a light CNN design consisting of 3 CL with
335 single-channel inputs Vy. This configuration enables the number of learnable parameters to be reduced
336 by a factor of 32 in comparison with the ResNet18. Moreover, we achieved a processing time of 24.6 ms
337 which is shorter than MEs (40 ms). Our study opens an interesting way to find the trade-off between
338 speed and accuracy in ME recognition. While the results are encouraging, it should be noted that our
339 method does not give a better accuracy than the ones described in the literature. Instead, a compromise
340 has to be made between accuracy and processing time. By minimizing the computation, our proposed
341 method manages to obtain accuracy comparable to the state-of-the-art systems while being compatible
342 with the real-time constraint of embedded vision.

343 Several future works could further enhance both the speed and accuracy of our proposed
344 ME recognition pipeline. These include more advanced data augmentation techniques to improve
345 recognition performance. Moreover, new ways to automatically optimize the structure of a network
346 to make it lighter have been presented recently. Other networks optimized for efficiency will also be
347 explored. For example, MobileNet [40] uses depth-wise separable convolutions to build light weight
348 CNN. ShuffleNet [41] uses pointwise group convolution to reduce computation complexity of 1x1
349 convolutions and channel shuffle to help the information flowing across feature channels. Our next
350 step of exploration aims to analyze and integrate these new methodologies in our framework.

351 **Author Contributions:** Conceptualization, C.M, F.Y; formal analysis, R.B, Y.L; methodology, R.B, Y.L, C.M;
352 software R.B, Y.L; supervision, C.M, D.G, F.Y; writing–review and editing, R.B, Y.L, C.M, D.G, F.Y; ;
353 Writing—original draft R.B, Y.L; funding acquisition, D.G. All authors have read and agreed to the published
354 version of the manuscript.

355 **Funding:** This work was supported by the H2020 Innovative Training Network (ITN) project ACHIEVE
356 (H2020-MSCA-ITN-2017: agreement no. 765866).

357 **Conflicts of Interest:** The authors declare no conflict of interest.

358 Abbreviations

359 The following abbreviations are used in this manuscript:

| | | |
|-----|----|---------------------|
| 360 | ME | Micro Expression |
| 361 | CL | Convolutional Layer |
| | M | Magnitude |

362 References

- 363 1. Shan, C.; Gong, S.; McOwan, P. Facial expression recognition based on local binary patterns: a
364 comprehensive study. *Image Vis. Comput.* **2018**, *27*, 803–816.
- 365 2. Edwards, J.; Jackson, H.; Pattison, P. Emotion recognition via facial expression and affective prosody in
366 schizophrenia: a methodological review. *Clin. Psychol. Rev.* **2002**, *22*, 789–832.
- 367 3. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S.
368 Analysis of emotion recognition using facial expressions, speech and multimodal information. *Int. Conf.*
369 *on Multimodal Interfaces* **2004**, pp. 205–211.
- 370 4. Ekman, P.; Friesen, W.V. Nonverbal Leakage and Clues to Deception. *Psychiatry* **1969**, *32*, 88–106.
- 371 5. Ekman, P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*; WW Norton
372 & Company, 2009.
- 373 6. Haggard, E.; Isaacs, K. Micromomentary facial expressions as indicators of ego mechanisms in
374 psychotherapy. *Methods of Research in Psychotherapy* **1966**, pp. 154–165.
- 375 7. Vecchiato, G.; Astolfi, L.; Fallani, F. On the use of EEG or MEG brain imaging tools in neuromarketing
376 research. *Comput. Intell. Neurosci.* **2011**.
- 377 8. Nass, C.; Jonsson, M.; Harris, H.; Reaves, B.; Endo, J.; Brave, S.; Takayama, L. Improving Automotive Safety
378 by Pairing Driver Emotion and Car Voice Emotion. *Extended Abstracts on Human Factors in Computing*
379 *Systems* **2005**, pp. 1973–1976.
- 380 9. Ekman, P. Lie Catching and Micro Expressions. *The Philosophy of Deception* **2009**, pp. 118–133.

- 381 10. Frank, M.; Herbasz, M.; Sinuk, K.; Keller, A.; Nolan, C. I see how you feel: training laypeople and
382 professionals to recognize fleeting emotions. *The Annual Meeting of International Communication Association*
383 **2009**.
- 384 11. Wang, S.J.; Yan, W.J.; Li, X.; Zhao, G.; Zhou, C.G.; Fu, X.; Yang, M.; Tao, J. Micro expression recognition
385 using color spaces. *Trans. on Image Process.* **2015**, *24*, 6034–6047.
- 386 12. Wu, Q.; Shen, X.; Fu, X. The Machine Knows What You Are Hiding: An Automatic Micro-Expression
387 Recognition System. *Affective Computing and Intelligent Interaction*, 2011, pp. 152–162.
- 388 13. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognising spontaneous facial micro-expressions. *ICCV* **2011**,
389 pp. 1449–1456.
- 390 14. Yan, W.; Li, X.; Wang, S.; Zhao, G.; Liu, Y.; Chen, Y.; Fu, X. CASMEII: an improved spontaneous
391 micro-expression database and the baseline evaluation. *PLOS One* **2014**, *9*, 1–8.
- 392 15. Davison, A.; Yap, M.; Costen, N.; Tan, K.; Lansley, C.; Leightley, D. Micro-facial movements: an
393 investigation on spatio-temporal descriptors. *ECCV* **2014**, pp. 111–123.
- 394 16. He, J.; Hu, J.F.; Lu, X.; Zheng, W.S. Multi-task mid-level feature learning for micro-expression recognition.
395 *Pattern Recognition* **2017**, *66*, 44–52.
- 396 17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *CVPR* **2016**, pp. 770–778.
- 397 18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *Int.*
398 *Conf. on Learning Representations* **2015**.
- 399 19. Liong, S.T.; See, J.; Phan, R.W.; Ngo, A.L.; Oh, Y.H.; Wong, K. Subtle expression recognition using optical
400 strain weighted features. *ACCV* **2014**.
- 401 20. Ruiz-Hernandez, J.; Pietikäinen, M. Encoding local binary patterns using re-parameterization of the second
402 order Gaussian jet. *Int. Conf. on Automatic Face and Gesture Recogn.* **2013**, pp. 1–6.
- 403 21. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A spontaneous micro-expression database:
404 Inducement, collection and baseline. *Int. Conf. on Aut. Face and Gesture Recogn.* **2013**, pp. 1–6.
- 405 22. Wang, Y.; See, J.; Phan, R.; Oh, Y. LBP with six intersection points: reducing redundant information in
406 LBP-TOP for micro-expression recognition. *Asian Conf. on Comp. Vis.* **2014**, pp. 525–537.
- 407 23. Huang, X.; Zhao, G.; Hong, X.; Zheng, W.; Pietikäinen, M. Spontaneous facial micro-expression analysis
408 using Spatiotemporal Completed Local Quantized Patterns. *Neurocomputing* **2016**, *175*, 564–578.
- 409 24. Liu, Y.J.; Zhang, J.K.; Yan, W.J.; Wang, S.J.; Zhao, G.; Fu, X. A Main directional mean optical flow feature
410 for spontaneous micro-expression recognition. *Trans. Affect. Comput.* **2015**, *7*, 299–310.
- 411 25. Oh, Y.H.; Ngo, A.C.L.; See, J.; Liong, S.T.; Phan, R.C.W.; Ling, H.C. Monogenic Riesz wavelet representation
412 for micro-expression recognition. *Int. Conf. on Digital Signal Proc.* **2015**, pp. 1237–1241.
- 413 26. Min, P.; Chongyang, W.; Tong, C.; Guangyuan, L.; Xiaolan, F. Dual Temporal Scale Convolutional Neural
414 Network for Micro-Expression Recognition. *Frontiers in Psychology* **2017**, *8*, 1745–1757.
- 415 27. Khor, H.Q.; See, J.; Phan, R.C.W.; Lin, W. Enriched Long-term Recurrent Convolutional Network for Facial
416 Micro-Expression Recognition. *Int. Conf. on Aut. Face Gesture Recogn.* **2018**, *1*, 667–674.
- 417 28. Liong, S.T.; Gan, Y.; See, J.; Khor, H.Q.; Huang, Y.C. Shallow Triple Stream Three-dimensional CNN
418 (STSTNet) for Micro-expression Recognition. *Int. Conf. on Aut. Face Gesture Recogn.* **2019**, pp. 1–5.
- 419 29. Patel, D.; Hong, X.; ; Zhao, G. Selective deep features for micro expression recognition. *Int. Conf. on*
420 *Pattern Recogn.* **2016**, pp. 2258–2263.
- 421 30. Li, Y.; Huang, X.; ; Zhao, G. Can micro-expression be recognized based on single apex frame? *Int. Conf. on*
422 *Image Proc.* **2018**, pp. 3094–3098.
- 423 31. Wang, S.J.; Li, B.J.; Liu, Y.J.; Yan, W.J.; Ou, X.; Huang, X.; Xu, F.; ; Fu, X. Micro-expression recognition
424 with small sample size by transferring long-term convolutional neural network. *Neurocomputing* **2018**,
425 *312*, 251–262.
- 426 32. Gan, Y.; Liong, S.T.; Yau, W.C.; Huang, Y.C.; ; Tan, L.K. Off-apexnet on micro-expression recognition
427 system. *Signal Proc.:Image Comm.* **2019**, *74*, 129–139.
- 428 33. Hui, T.W.; Tang, X.; Loy, C.C. LiteFlowNet: A Lightweight Convolutional Neural Network for Optical
429 Flow Estimation. *Comp. Vis. and Pattern Analysis* **2018**.
- 430 34. Rieger, I.; Hauenstein, T.; Hettenkofer, S.; Garbas, J.U. Towards Real-Time Head Pose Estimation: Exploring
431 Parameter-Reduced Residual Networks on In-the-wild Datasets. *Int. Conf. on Industrial, Engineering and*
432 *Other Applications of Applied Intelligent Systems* **2019**, pp. 122–134.
- 433 35. Horn, B.; Schunck, B. Determining optical flow. *Artificial Intelligence* **1981**, *17*, 185–203.

- 434 36. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. SAMM: A Spontaneous Micro-Facial Movement
435 Dataset. *Trans. on Affective Comp.* **2018**, *9*, 116–129.
- 436 37. Davison, A.K.; Merghani, W.; Yap, M.H. Objective classes for micro-facial expression recognition. *J.*
437 *Imaging* **2018**, *4*.
- 438 38. Yap, M.H.; See, J.; Hong, X.; Wang, S.J. Facial Micro-Expressions Grand Challenge 2018 Summary. *Int.*
439 *Conf. On Aut. Face and Gesture Recogn.* **2018**, pp. 675–678.
- 440 39. Peng, M.; Wu, Z.; Zhang, Z.; Chen, T. From Macro to Micro Expression Recognition: Deep Learning on
441 Small Datasets Using Transfer Learning. *Int. Conf. on Aut. Face and Gesture Recogn.* **2018**, pp. 657–661.
- 442 40. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Andreetto, T.W.M.; Adam, H. MobileNets:
443 Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR* **2017**.
- 444 41. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for
445 Mobile Devices. *Comp. Vis. and Pattern Recogn.* **2018**, pp. 6848–6856.

446 **Sample Availability:** Samples of the compounds are available from the authors.

447 © 2020 by the authors. Submitted to *Appl. Sci.* for possible open access publication under the terms and conditions
448 of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).