applied sciences

MDPI

Article

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

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

ImViA EA 7535, University Bourgogne Franche-Comté, 21000 Dijon, France; Yu_liu@etu.u-Bourgogne.fr (Y.L.); Cyrille.Migniot@u-bourgogne.fr (C.M.); Dominique.Ginhac@u-bourgogne.fr (D.G.); fanyang@u-bourgogne.fr (F.Y.)

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

Received: 5 June 2020; Accepted: 16 July 2020; Published: 19 July 2020

check for updates

Abstract: Micro-Expression (ME) recognition is a hot topic in computer vision as it presents a gateway to capture and understand daily human emotions. It is nonetheless a challenging problem due to 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 intense in terms of memory consumption and computational resources. This poses great challenges when deploying CNN-based solutions in many applications, such as driver monitoring and comprehension recognition in virtual classrooms, which demand fast and accurate recognition. As these networks were 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 optimized for ME classification in two ways. Firstly, we reduced the depth of the network by removing residual layers. Secondly, we introduced 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 was 60.17% on the challenging composite dataset containing five 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 [1–4]. Most of the affective computing methods in the literature apply the emotion model presented by Ekman [5] that reported seven basic expressions: anger, fear, surprise, sadness, disgust, contempt and happiness. Ekman developed the Facial Action Coding System (FACS) to describe the facial muscle movements according to the action units , i.e., the fundamental actions of individual muscles or groups of muscles that can be combined to represent each of the facial expressions. These facial expressions can thus be labeled by codes based on the observed facial movements rather than from subjective classifications of emotion.

In contrast to the traditional macro-expression, people are less familiar with micro facial expressions [5,6], and even fewer know how to capture and recognize them. A Micro-Expression (ME) is a rapid and involuntary facial expression that exposes a person's true emotion [7]. 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 one deliberately prevents oneself from expressing genuine emotions. On the 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 the security field, neuromarketing [8], automobile drivers' monitoring [9] and lies and deceit detection [6].

Psychological research shows that facial MEs generally are transient (e.g., remaining less than 200 ms) and very subtle [10]. The short duration and subtlety levy great challenges on a human trying to perceive and recognize them. To enable better ME recognition by humans, 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 [11]. Moreover, humans' decisions are prone to being influenced by individual perceptions that vary among subjects and across 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 have 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 textures of faces. Most existing methods [12,13] attempt to extract low-level features, such as the widely used Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) [14–16] 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 for inactive facial regions that are less dynamic [17].

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 [18] and VGG [19]. These studies re-purposed the use of off-the-shelf CNNs by giving them input data taken from the optical flow extracted from the MEs. While achieving good performance, these neural networks are quite demanding in terms of memory usage and computation.

In specific applications, for example, during automobile driver monitoring or student 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 was to design an efficient and accurate ME recognition pipeline for embedded vision purposes. First of all, our design took into account thorough investigations on different CNN architectures. Next, different optical flow representations for CNN inputs were studied. Finally, our proposed pipeline achieved accuracy for ME recognition that is competitive with 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 studies 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 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 patters between image pixels, and the temporal relationship between pixels and their neighboring frames. The resulting histograms are then concatenated to represent the temporal changes over entire videos. LBP-TOP has been widely adopted in several studies. Pfister et al. [14] applied LBP-TOP for spontaneous ME recognition. Yan et al. [15] achieved 63% ME recognition accuracy on their CASME II database using LBP-TOP. In addition, LBP-TOP has also been used to investigate differences between micro-facial movement sequences and neutral face sequences.

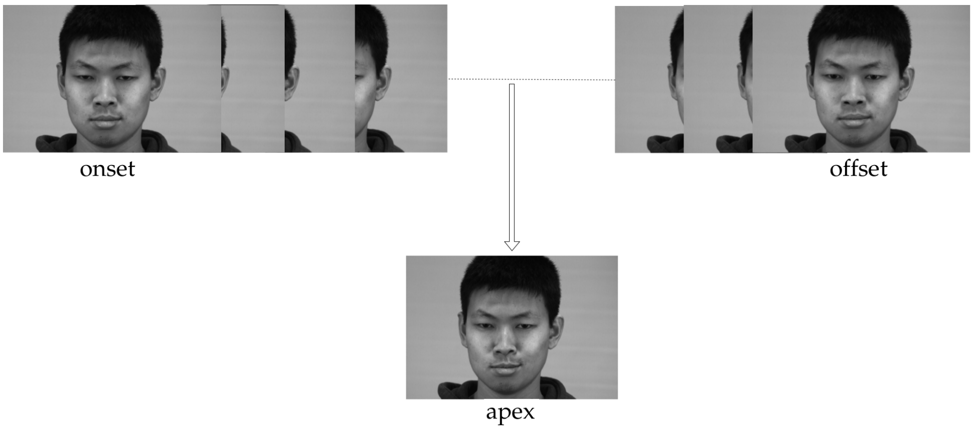


Figure 1. Example of a Micro-Expression (ME): the maximum movement intensity occurs at the apex frame.

Several studies aimed to extend low-level features extracted by LBP-TOP as they still could not reach satisfactory accuracy. For example, Liong et al. [20] proposed assigning different weights to local features, thereby putting more attention on active facial regions. Wang et al. [12] studied the correlation between color and emotions by extracting LBP-TOP from the Tensor-Independent Color Space (TICS). Ruiz-Hernandez and Pietikäinen [21] used the re-parameterization of second order Gaussian jet on the LBP-TOP, achieving a promising ME recognition result on the SMIC database [22]. Considering that LBP-TOP consists of redundant information, Wang et al. [23] proposed the LBP-Six Intersection Points (LBP-SIP) method which is computationally more efficient and achieves higher accuracy on the CASEME II database. We also note that the STCLQP (Spatio-Temporal Completed Local Quantization Patterns) proposed by Huang et al. [24] achieved a substantial improvement for analyzing facial MEs.

Over the years, as research showed that it is non-trivial for low-level features to effectively capture and encode a ME's subtle dynamic patterns (especially from inactivate regions), other methods shifted to exploit mid or high-level features. He et al. [17] developed a novel multi-task mid-level feature learning method to enhance the discrimination ability of the extracted low-level features. The mid-level feature representation is generated by learning a set of class-specific feature mappings. Better recognition performance has been obtained with more available information and features more suited to discrimination and generalization. A simple and efficient method known as Main Directional Mean Optical-flow (MDMO) was employed by Liu et al. [25]. They used optical flow to measure the subtle movement of facial Regions of Interest (ROIs) that were spotted based on the FACS. Oh et al. [26] also applied the monogenic Riesz wavelet representation in order to amplify subtle movements of MEs.

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

Peng et al. [27] proposed a two-stream convolutional network DTSCNN (Dual Temporal Scale Convolutional Neural Network) to address two aspects: the overfitting problem caused by the small sizes of existing ME databases and the use of high-level features. We can observe four characteristics of the DTSCNN: (i) separate features were first extracted from ME clips from two shallow networks and then fused; (ii) data augmentation and higher drop-out ratio were applied in each network; (iii) two databases (CASME I and CASME II) were combined to train the network; (iv) the data fed to the networks were optical-flow images instead of raw RGB frames.

Khor et al. [28] studied two variants of an Enriched LRCN (Long-Term Recurrent Convolutional Network) model for ME recognition. Spatial Enrichment (SE) refers to channel-wise stacking of gray-scale and optical flow images as new inputs to CNN. On the other hand, Temporal Enrichment (TE) stacks obtained features. Their TE model achieves better accuracy on a single database, while the SE model is more robust against the cross-domain protocol involving more databases.

Liong et al. [29] designed a Shallow Triple Stream Three-dimentional CNN (STSTNet). The model takes input stacked optical flow images computed between the onset and apex frames (optical strain, horizontal and vertical flow fields), followed by three shallow Convolutional Layers in parallel and a fusion layer. The proposed method is able to extract rich features from MEs while being computationally light, as the fused features are compact yet discriminative.

Our objective was to realize a fast and high-performance ME recognition pipeline for embedded vision applications under several constraints, such as embeddability, limited memory and restricted computing resources. Inspired by existing works [27,29], we explored different CNN architectures and several optical flow representations for CNN inputs to find cost-effective neural network architectures that were capable of recognizing MEs in real-time.

3. Methodology

The studies applying deep learning to tackle the ME classification problem [30–33] usually used pretrained CNNs such as ResNet [18] and VGG [19] and applied transfer learning to obtain ME features. In our work, we first selected off-the-shelf ResNet18 because it provided the best trade-off between accuracy and speed on the challenging ImageNet classification and was recognized for its performance in transfer learning. ResNet [18] explicitly lets the stacked layers fit a residual mapping. Namely, the stacked non-linear layers are let to fit another mapping of $F(x) := H(x) - x$ where $H(x)$ is the desired underlying mapping and x the initial activations. The original mapping is recast into $F(x) + x$ by feedforward neural networks with shortcut connections. ResNet18 has 20 Convolutional Layers (CLs) (17 successive CLs and 3 branching ones). Residual links after each pair of successive convolutional units are used and the kernel size after each residual link is doubled. As ResNet18 is designed to extract features from RGB color images, it requires inputs to have 3 channels.

In order to accelerate processing speed in the deep learning domain, the main current trend in decreasing complexity of CNN is to reduce the number of parameters. For example, Hui et al. [34] proposed a very compact LiteFlowNet which is 30 times smaller in the model size and 1.36 times faster in the running speed in comparison with the state-of-the-art CNNs for optical flow estimation. In [35], Rieger et al. explored parameter-reduced residual networks on in-the-wild datasets, targeting real-time head pose estimation. They experimented with various ResNet architectures with a varying number of layers to handle different image sizes (including low-resolution images). The optimized ResNet achieved state-of-the-art accuracy with real-time speed.

It is well known that CNNs are created for specific problems and therefore over-calibrated when they are used in other contexts. ResNet18 was made for end-to-end object recognition: the dataset used for training had hundreds of thousands of images for each class and more than a thousand classes in total. Based on that: (i) An ME recognition study considers at most 5 classes, and the datasets of spontaneous MEs are scarce and contain far fewer samples, and (ii) optical flows are high-level features contrary to low-level color features and so require shallower networks; we have empirically reduced the architecture of ResNet18 by iteratively removing residual layers. This allowed us to assess the influence of the depth of the network on its classification capacities in our context and therefore to estimate the relevant calibration of the network.

Figure 2 illustrates the reduction protocol: at each step the last residual layer with two CLs is removed and the previous one is connected to the fully connected layer. Only networks with an odd number of CL are therefore proposed. The weights of all CNNs are pretrained using ImageNet. As highlighted in Table 1, the decrease in the number of CLs has a significant impact on the number of learnable parameters of the network, which directly affects the forward propagation time.

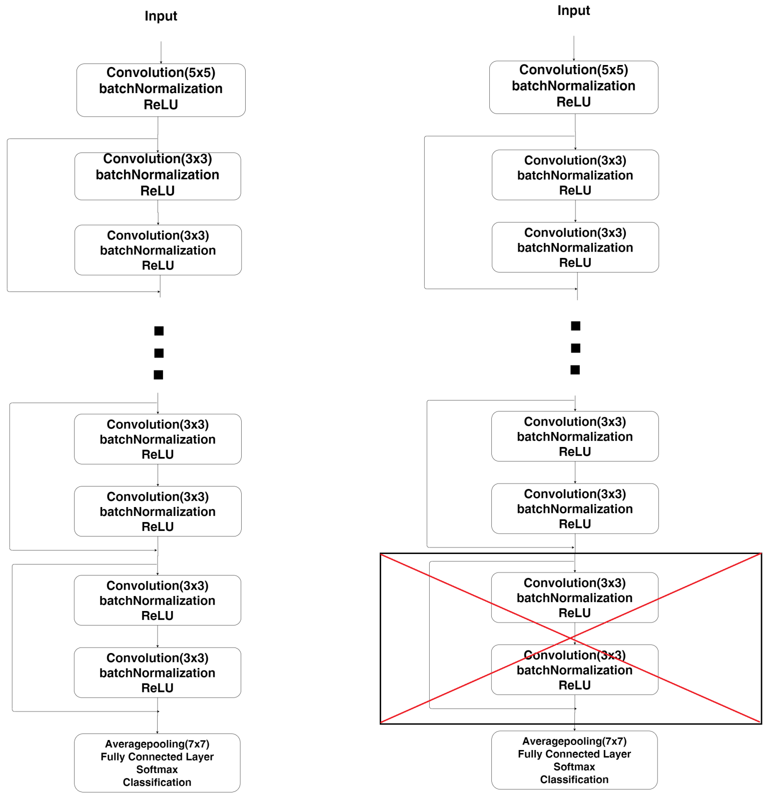


Figure 2. Depth reduction of a deep neural network: in the initial network, each residual layer contains two Convolutional Layers (CLs) (**left**); the last residual layer is removed (**middle**) to obtain a shallower network (**right**).

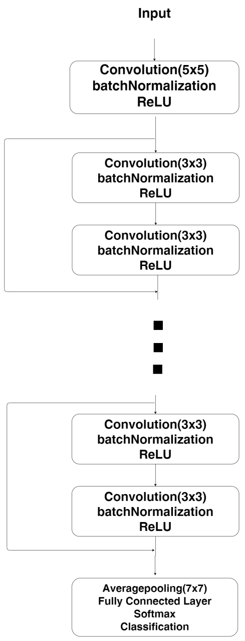


Table 1. Number of CLs and the number of learnable parameters in the proposed architectures.

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

Once the network depth has been correctly estimated, the dimensions of the input have to be optimized. In our case, CNNs take optical flows extracted between the onset and apex frames of

ME video clips. It is between these two moments that the motion is most likely to be the strongest. The dimensionality of inputs determines the complexity of the network that uses them, since the reduction in input channels dictates the number of filters to be used throughout all following layers of the CNN. The optical flow between the onset (Figure 3a) and the apex (Figure 3b) typically has a 3-channel representation to be used in a pretrained architecture designed for 3-channel color images. This representation, however, may not be optimal for ME recognition.

From the assumption of brightness invariance, the movement of each pixel between frames over a period of time is estimated and represented as a vector (Figure 3c) indicating the direction and intensity of the motion. The projection of the vector on the horizontal axis corresponds to the Vx field (Figure 3d) while its projection on the vertical axis is the Vy field (Figure 3e). The Magnitude (M) is the norm of the vector (Figure 3f). Figure 4 illustrates this representation of one optical flow vector.

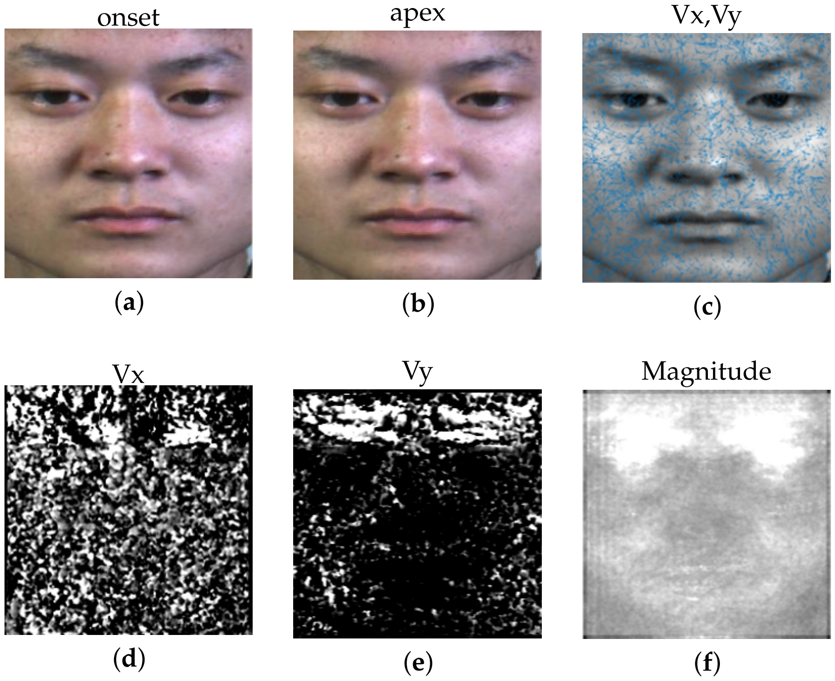


Figure 3. Optical flow is computed between the onset (**a**) and the apex (**b**): vectors obtained for a random sample of pixels (**c**), Vx field (**d**), Vy field (**e**) and Magnitude field (**f**).

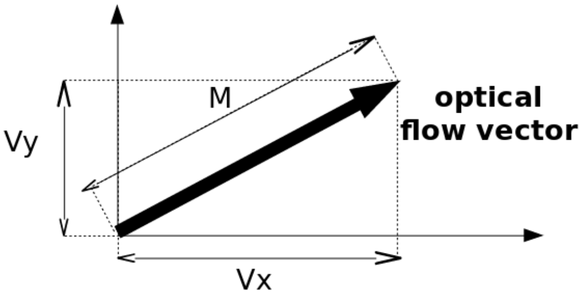


Figure 4. Visualization of M, Vx and Vy for one optical flow vector.

When classifying ME, the resulting matrices Vx , Vy and M are traditionally given as inputs to the CNN. Nonetheless, the third channel is inherently redundant since M is computed from Vx and Vy. Optical flow composed of the 2-channel Vx and Vy field could already provide all relevant information. Furthermore, we hypothesize that even a single channel motion field itself could be descriptive enough. Hence, we have created and evaluated networks taken as input for the optical flow in a two-channel representation (Vx-Vy) and in an one-channel representation (M, Vx or Vy).

For this purpose, the proposed networks begin with a number of CLs related to the depth optimization followed by a batch normalization and ReLU. Then the networks end with a maxpooling layer and a fully connected layer. The Figure 5 presents the architectures used with one to four CL according to the results of the experiments in Section 4. As illustrated in Table 2, a low dimensional input leads to a significant reduction in the number of learnable parameters and therefore in the complexity of the system.

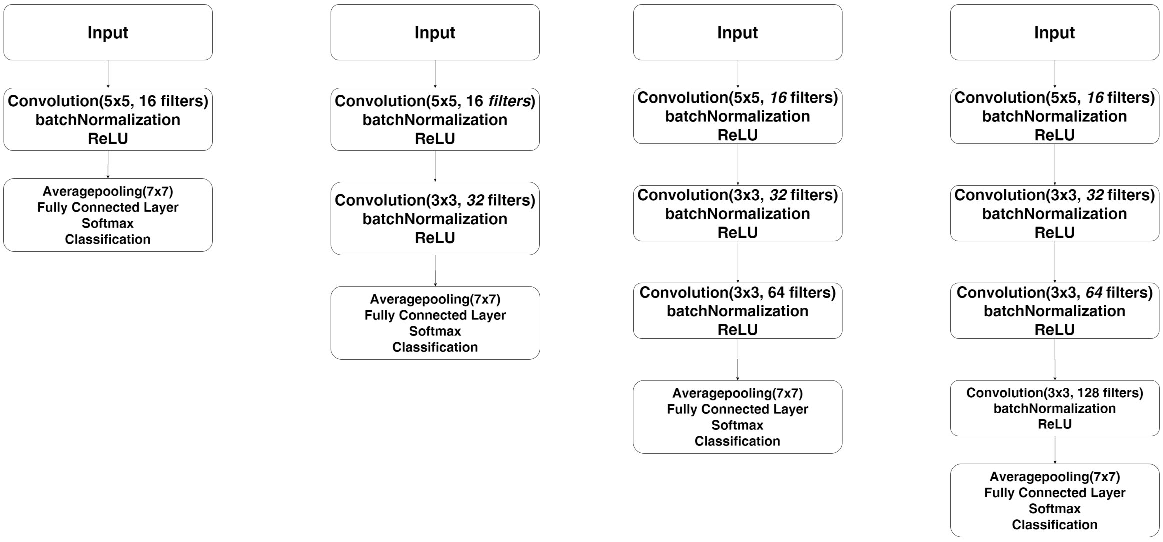

Input
Convolution(5x5, 16 filters) batchNormalization ReLU
Averagepooling(7x7) Fully Connected Layer Softmax Classification
Input
Convolution(5x5, 16 filters) batchNormalization ReLU
Convolution(3x3, 32 filters) batchNormalization ReLU
Averagepooling(7x7) Fully Connected Layer Softmax Classification
Input
Convolution(5x5, 16 filters) batchNormalization ReLU
Convolution(3x3, 32 filters) batchNormalization ReLU
Convolution(3x3, 64 filters) batchNormalization ReLU
Averagepooling(7x7) Fully Connected Layer Softmax Classification
Input
Convolution(5x5, 16 filters) batchNormalization ReLU
Convolution(3x3, 32 filters) batchNormalization ReLU
Convolution(3x3, 64 filters) batchNormalization ReLU
Convolution(3x3, 128 filters) batchNormalization ReLU
Averagepooling(7x7) Fully Connected Layer Softmax Classification

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

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

4. Experiments

4.1. Dataset and Validation Protocol Presentation

Two ME databases were used in our experiments. CASME II (Chinese Academy of Sciences Micro-Expression) [15] is a comprehensive spontaneous ME database containing 247 video samples, collected from 26 Asian participants with an average age of 22.03 years old. Compared to the first database, the Spontaneous Actions and Micro-Movements (SAMM) [36] is a more recent one consisting of 159 micro-movements (one video for each). These videos were collected spontaneously from a demographically diverse group of 32 participants with a mean age of 33.24 years old and a balanced gender split. Originally intended for investigating micro-facial movements, SAMM initially collected the seven basic emotions.

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

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

The Horn–Schunck method [39] was selected to compute optical flow. This algorithm was widely used for optical flow estimation in many recent studies for virtue of its robustness and efficiency. Throughout all experiments, we trained the CNN models with a mini-batch size of 64 for 150 epochs using the RMSprop optimization. Feature extraction and classification were both handled by the CNN. Simple data augmentation was applied to double the training size. Specifically, for each ME video clip used for training, in addition to the optical flow between the onset and apex frame, we also included a second flow computed between the onset and apex+1 frame.

4.2. ResNet Depth Study

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

Table 3. Accuracies varied by the number of Convolutional Layers (CLs) 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%

We observed that the best score was achieved by ResNet8, which had seven CLs. However, the scores achieved by different numbers of CL did not vary much. Furthermore, beyond seven CL, adding more CL did not improve the accuracy of the model. The fact that accuracy does not increase along with depth confirms that multiple successive CL are not necessary to achieve a respectable accuracy. The most interesting observation was that with a single CL, we achieved a score that is not very far from the optimal score while the size of the model was much more concise. This suggests that instead of deep learning, a more "classical" approach exploiting shallow neural networks presents an interesting field to explore when considering portability and computational efficiency for embedded systems. That is the principal reason we restricted our study to shallow CNNs.

4.3. CNN Input Study

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

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

	1 CL	**2 CL**	**3 CL**	**4 CL**
Vx	52.24%	54.34%	53.92%	53.50%
Vy	58.09%	59.34%	**60.17%**	60.17%
Vx-Vy	58.51%	59.75%	**60.17%**	58.09%
M	58.09%	58.92%	59.34%	59.34%

We can observe that the Vx-Vy pair and Vy alone gave the best results, both representations achieving 60.17% accuracy. On the other hand, using Magnitude alone leads to a similar accuracy to those of Vy and the Vx-Vy pair with a score of 59.34%. Vx got the worst results overall, with a maximum score of 54.34%. This observation indicates that the most prominent features for ME classification might indeed be more dominant in vertical movement rather than the horizontal movement. This assumption is logical when thinking about the muscle movements happening in each known facial expression.

To better visualize the difference in the high-level features present in Vx, Vy and the Magnitude, we did an averaging on all the different samples according to their classes. The result can be seen in Figure 6. We observed that Vx exhibits a non-negligible quantity of noise. Magnitude and Vy, on the other hand, had clear regions of activity for each class. The regions of activity were aligned with the muscles responsible of each facial expression.

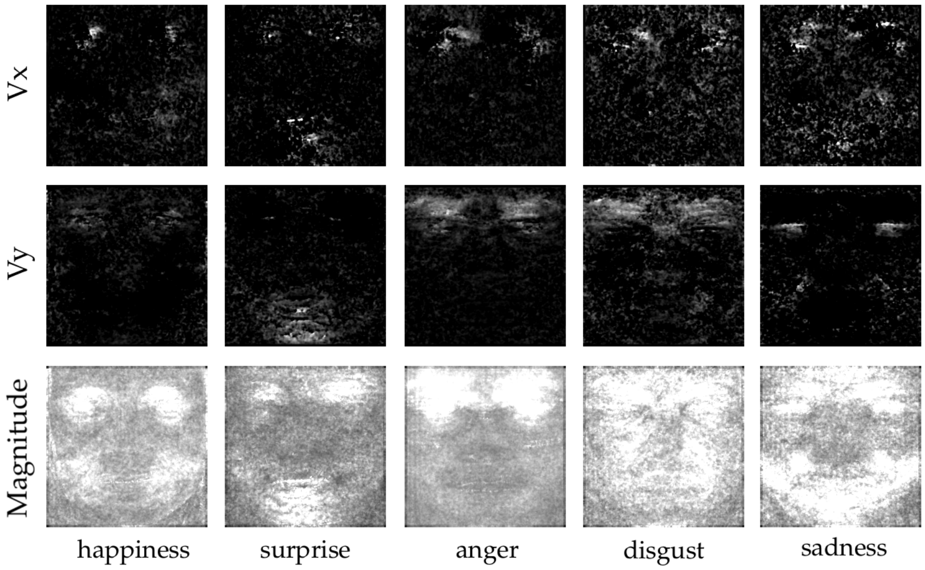


Figure 6. Average optical flow obtained in the dataset per ME class. Studied classes are in order from left to right: happiness, surprise, anger, disgust and sadness.

4.4. Classification Analysis

In order to understand obtained results, we measured cosine similarity of features extracted by three CNNs: ResNet8 (Section 4.2), Vx-Vy-3-CL and Vy-3-CL (Section 4.3). Usually, the convolutional layers of CNNs are considered as different feature extractors; only the last fully connected layer directly performs the classification task. The features just before classification can be represented in vector format. Cosine similarity measures the similarity between two vectors a and b using Equation (1):

$$cosine(a,b) = \frac{a^T b}{||a||\ ||b||} \tag{1}$$

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

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

	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 6. Cosine similarity for the 3-CL CNN with double-channel inputs (Vx-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 7. Cosine similarity for ResNet8.

	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

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

4.5. Performance Evaluations

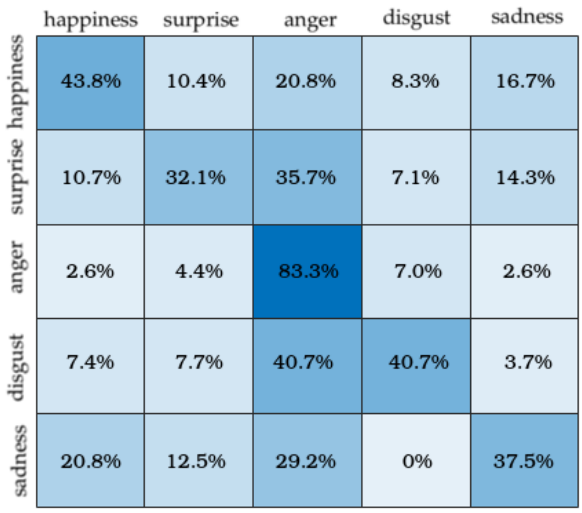
In this subsection, we describe measuring our proposed method in three aspects: recognition accuracy, needed memory space and processing speed. Since we obtained optimal results by using the Vy field and 3-layer CNN, further evaluations concentrated on this particular configuration.

Evaluation on recognition accuracy: We performed an accuracy comparison of five objective ME class recognition (see Table 8). Our best CNN reached a similar performance as those of other studies using the same protocol of validation. It is worth mentioning that Peng et al. [40] employed a macro-to-micro transferred ResNet10 model and obtained a better result. Their work used four Macro-Expression datasets (>10 K images) and some preprocessing, such as color shift, rotation and smoothing. These additional operations make their proposed method difficult for deployment on embedded systems. After seeing the confusion matrix of our model (Figure 7), we also noticed that the distribution of correct assessments for Vy was more balanced than the ones gotten from [28] (Figure 8).

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

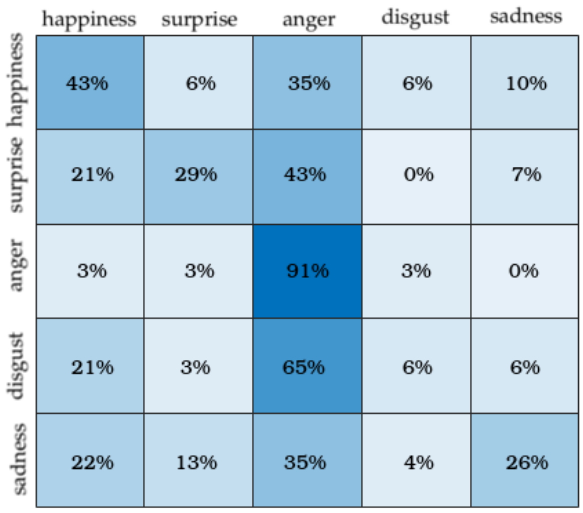
Method	**Accuracy**
LBP_TOP [28]	42.29%
Khor et al. [28]	57.00%
Peng et al. [40]	74.70%
Proposed method	60.17%

The DTSCNN proposed by Peng et al. in [27] opted for two optical flows computed differently from a ME sample, which made the whole network robust to different frame rates of ME videos. In detail, the first optical flow is calculated using 64 frames around the apex to adapt to the frame rate of CASME I. Similarly, the second optical flow is given by the 128 frames around the apex adapted to the frame rate of CASME II. In case the number of frames composing the ME is not sufficient, a linear interpolation method is used to normalize the video clips. Their study used two CNNs in parallel to extract two separate features before concatenating them. The resultant feature vector was then fed as input to an SVM to be classified. The DTSCNN was tested on four classes (positive, negative, surprise and other) from a composite dataset consisting of the CASME I and CASME II databases, and it achieved an average recognition rate of 66.67%. The STSTNet proposed by Liong et al. in [29] makes use of three-dimensional CNNs which carry out three-dimensional convolutions instead of two-dimensional ones (such as ResNet, VGG, the networks presented in [27,28,40] and our study). It was tested on three classes: positive, negative and surprise from a composite database consisting of samples from the SMIC, CASME II and SAMM databases. It achieved an unweighted average recall rate of 76.05% and an unweighted F1-score of 73.53%. Both of these two frameworks are not very suitable for real-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 CLs and Vy as input.



	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 [28].

Evaluation on memory space: Table 9 summarizes the number of learnable parameters and used filters according to the dimensionality of the network inputs. The minimum required memory space 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)

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

5. Conclusions and Future Works

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

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

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

Several future works could further enhance both the speed and accuracy of our proposed ME recognition pipeline. These include more advanced data augmentation techniques to improve recognition performance. Moreover, new ways to automatically optimize the structure of a network to make it lighter have been presented recently. Other networks optimized for efficiency will also be explored. For example, MobileNet [41] uses depth-wise separable convolutions to build light weight CNN. ShuffleNet [42] uses pointwise group convolution to reduce computation complexity of 1 × 1 convolutions and channel shuffle to help the information flowing across feature channels. Our next step of exploration aims to analyze and integrate these new methodologies in our framework. Furthermore, we also hope to investigate new emotional machines while avoiding AI modeling errors and biases [43].

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

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

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

Abbreviations

The following abbreviations are used in this manuscript:

ME	Micro Expression
CL	Convolutional Layer
M	Magnitude

References

1. Shan, C.; Gong, S.; McOwan, P. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2018**, *27*, 803–816. [CrossRef]
2. Edwards, J.; Jackson, H.; Pattison, P. Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review. *Clin. Psychol. Rev.* **2002**, *22*, 789–832. [CrossRef]
3. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces, College, PA, USA, 13–15 October 2004; pp. 205–211.
4. Biondi, G.; Franzoni, V.; Gervasi, O.; Perri, D. An Approach for Improving Automatic Mouth Emotion Recognition. In *Lecture Notes in Computer Science, 11619 LNCS*; Springer, Cham: Switzerland, 2019; pp. 649–664.
5. Ekman, P.; Friesen, W.V. Nonverbal Leakage and Clues to Deception. *Psychiatry* **1969**, *32*, 88–106. [CrossRef] [PubMed]
6. Ekman, P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*; WW Norton & Company: New York, NY, USA, 2009.
7. Haggard, E.; Isaacs, K. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of Research in Psychotherapy*; Springer: Boston, MA, USA, 1966; pp. 154–165.
8. Vecchiato, G.; Astolfi, L.; Fallani, F. On the use of EEG or MEG brain imaging tools in neuromarketing research. *Comput. Intell. Neurosci.* 2011. [CrossRef] [PubMed]
9. Nass, C.; Jonsson, M.; Harris, H.; Reaves, B.; Endo, J.; Brave, S.; Takayama, L. Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In Proceedings of the Extended Abstracts on Human Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, 2005; pp. 1973–1976.
10. Ekman, P. Lie Catching and Micro Expressions. In *The Philosophy of Deception*; Oxford University Press : Oxford, UK, 2009; pp. 118–133.
11. Frank, M.; Herbasz, M.; Sinuk, K.; Keller, A.; Nolan, C. I see how you feel: training laypeople and professionals to recognize fleeting emotions. In Proceedings of the Annual Meeting of International Communication Association, Chicago, IL USA, 21–25 May 2009.
12. Wang, S.J.; Yan, W.J.; Li, X.; Zhao, G.; Zhou, C.G.; Fu, X.; Yang, M.; Tao, J. Micro expression recognition using color spaces. *Trans. Image Process.* **2015**, *24*, 6034–6047. [CrossRef] [PubMed]
13. Wu, Q.; Shen, X.; Fu, X. The Machine Knows What You Are Hiding: An Automatic Micro-Expression Recognition System. In Proceedings of the Affective Computing and Intelligent Interaction, Memphis, TN, USA, 9–12 October 2011; pp. 152–162.
14. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognising spontaneous facial micro- expressions. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1449–1456.
15. Yan, W.; Li, X.; Wang, S.; Zhao, G.; Liu, Y.; Chen, Y.; Fu, X. CASMEII: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **2014**, *9*, e86041.

16. Davison, A.; Yap, M.; Costen, N.; Tan, K.; Lansley, C.; Leightley, D. Micro-facial movements: an investigation on spatio-temporal descriptors. In Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2014; pp. 111–123.
17. He, J.; Hu, J.F.; Lu, X.; Zheng, W.S. Multi-task mid-level feature learning for micro-expression recognition. *Pattern Recognit.* **2017**, *66*, 44–52. [CrossRef]
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
20. Liong, S.T.; See, J.; Phan, R.W.; Ngo, A.L.; Oh, Y.H.; Wong, K. Subtle expression recognition using optical strain weighted features. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014.
21. Ruiz-Hernandez, J.; Pietikäinen, M. Encoding local binary patterns using re-parameterization of the second order Gaussian jet. In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–6.
22. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–6.
23. Wang, Y.; See, J.; Phan, R.; Oh, Y. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 525–537.
24. Huang, X.; Zhao, G.; Hong, X.; Zheng, W.; Pietikäinen, M. Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns. *Neurocomputing* **2016**, *175*, 564–578. [CrossRef]
25. Liu, Y.J.; Zhang, J.K.; Yan, W.J.; Wang, S.J.; Zhao, G.; Fu, X. A Main directional mean optical flow feature for spontaneous micro-expression recognition. *Trans. Affect. Comput.* **2015**, *7*, 299–310. [CrossRef]
26. Oh, Y.H.; Ngo, A.C.L.; See, J.; Liong, S.T.; Phan, R.C.W.; Ling, H.C. Monogenic Riesz wavelet representation for micro-expression recognition. In Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 21–24 July 2015; pp. 1237–1241.
27. Min, P.; Chongyang, W.; Tong, C.; Guangyuan, L.; Xiaolan, F. Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition. *Front. Psychol.* **2017**, *8*, 1745–1757.
28. Khor, H.Q.; See, J.; Phan, R.C.W.; Lin, W. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition. In Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; Volume 1, pp. 667–674.
29. Liong, S.T.; Gan, Y.; See, J.; Khor, H.Q.; Huang, Y.C. Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–5.
30. Patel, D.; Hong, X.; Zhao, G. Selective deep features for micro expression recognition. In Proceedings of the International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2258–2263.
31. Li, Y.; Huang, X.; Zhao, G. Can micro-expression be recognized based on single apex frame? In Proceedings of the International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3094–3098.
32. Wang, S.J.; Li, B.J.; Liu, Y.J.; Yan, W.J.; Ou, X.; Huang, X.; Xu, F.; Fu., X. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* **2018**, *312*, 251–262. [CrossRef]
33. Gan, Y.; Liong, S.T.; Yau, W.C.; Huang, Y.C.; Tan., L.K. Off-apexnet on micro-expression recognition system. *Signal Proc. Image Comm.* **2019**, *74*, 129–139. [CrossRef]
34. Hui, T.W.; Tang, X.; Loy, C.C. LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
35. Rieger, I.; Hauenstein, T.; Hettenkofer, S.; Garbas, J.U. Towards Real-Time Head Pose Estimation: Exploring Parameter-Reduced Residual Networks on In-the-wild Datasets. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Graz, Austria, 9–11 July 2019; pp. 122–134.

36. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. SAMM: A Spontaneous Micro-Facial Movement Dataset. *Trans. Affective Comp.* **2018**, *9*, 116–129. [CrossRef]
37. Davison, A.K.; Merghani, W.; Yap, M.H. Objective classes for micro-facial expression recognition. *J. Imaging* **2018**, *4*, 119. [CrossRef]
38. Yap, M.H.; See, J.; Hong, X.; Wang, S.J. Facial Micro-Expressions Grand Challenge 2018 Summary. In Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 675–678.
39. Horn, B.; Schunck, B. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]
40. Peng, M.; Wu, Z.; Zhang, Z.; Chen, T. From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning. In Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 657–661.
41. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Andreetto, T.W.M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
42. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
43. Vallverdù, J.; Franzoni, V. Errors, biases and overconfidence in artificial emotional modeling. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence Workshops, Thessaloniki, Greece, 14–17 October 2019; pp. 86–90.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).