sensors

MDPI

Article

Micro-Expression Recognition Based on Optical Flow and PCANet+

Shiqi Wang [1], Suen Guan [1], Hui Lin [1], Jianming Huang [2], Fei Long [1,2,*] and Junfeng Yao [1,2]

[1] School of Informatics, Xiamen University, Xiamen 361005, China; wangshiqi@stu.xmu.edu.cn (S.W.); gse2020XMU@foxmail.com (S.G.); 24320182203231@stu.xmu.edu.cn (H.L.); yao0010@xmu.edu.cn (J.Y.)
[2] Center for Digital Media Computing and Software Engineering, Xiamen University, Xiamen 361005, China; 24320142202428@stu.xmu.edu.cn
* Correspondence: flong@xmu.edu.cn

Abstract: Micro-expressions are rapid and subtle facial movements. Different from ordinary facial expressions in our daily life, micro-expressions are very difficult to detect and recognize. In recent years, due to a wide range of potential applications in many domains, micro-expression recognition has aroused extensive attention from computer vision. Because available micro-expression datasets are very small, deep neural network models with a huge number of parameters are prone to overfitting. In this article, we propose an OF-PCANet+ method for micro-expression recognition, in which we design a spatiotemporal feature learning strategy based on shallow PCANet+ model, and we incorporate optical flow sequence stacking with the PCANet+ network to learn discriminative spatiotemporal features. We conduct comprehensive experiments on publicly available SMIC and CASME2 datasets. The results show that our lightweight model obviously outperforms popular handcrafted methods and also achieves comparable performances with deep learning based methods, such as 3D-FCNN and ELRCN.

Keywords: micro-expression recognition; optical flow; PCANet+; deep learning

Citation: Wang, S.; Guan, S.; Lin, H.; Huang, J.; Long, F.; Yao, J. Micro-Expression Recognition Based on Optical Flow and PCANet+. *Sensors* **2022**, 22, 4296. https://doi.org/10.3390/s22114296

Academic Editors: Mehmet Rasit Yuce, Jan Cornelis and Christophoros Nikou

Received: 31 December 2021
Accepted: 31 May 2022
Published: 5 June 2022

1. Introduction

Micro-expressions (MEs) are involuntary facial movements with the characteristics of short duration, low intensity, and occurrence in sparse facial action units [1,2]. It is generally believed that the duration of ME is between 1/25 s and 1/2 s [3]. Micro-expression (ME) recognition is a challenging task; even the recognition accuracy by people with specialized training is below 50% [4,5]. Because MEs can reveal genuine emotions people try to hide [1,6], ME recognition has many potential applications in different fields, such as criminal investigation, commercial negotiation, clinical diagnosis, and so on [7,8]. Due to the characteristics of short duration and subtlety, how to extract discriminatory features from ME video clips is a key problem in the task of ME recognition [9]. In recent years, automatic detection and recognition of MEs has become an active research topic in computer vision [10–12].

In 2011, Pfister et al. [13] applied LBP-TOP (local binary pattern with three orthogonal planes) [14] to extract dynamic features of MEs on SMIC [12] dataset, and they proposed a benchmark framework for automatic ME recognition. In 2014, Yan et al. [15] established a new ME dataset called CASME2 and used LBP-TOP for ME recognition. Huang et al. [16] proposed a completed local quantization patterns (CLQP) method, which extends LQP by using the sign-based difference, the magnitude-based difference, and the orientation-based difference, and then converts them into binary codes. Wang et al. [17] proposed LBP with six intersection points (LBP-SIP) to obtain a more compact feature representation. The STLBP-IP [18] method proposed by Huang et al. uses integral projection based on difference image and LBP to extract the spatiotemporal features of MEs. In addition, Zong et al. [19]

expanded the effectiveness of the LBP Operator by layered STLBP-IP features and reduced the dimension of features by using the sparse learning method.

Lu et al. [20] proposed a Delaunay-based temporal coding model (DTCM) to represent spatiotemporally important features for MEs. Xu et al. [21] proposed a method called Facial Dynamic Map (FDM) to represent the movement patterns of MEs based on dense optical flow. Liu et al. [22] proposed a ME recognition method called Main Directional Mean Optical flow (MDMO), in which a face image is divided into 36 subregions, and the principal direction optical flow of all regions is connected to obtain a low dimensional feature vector. Liong et al. [23] proposed a method of ME detection and recognition by using optical strain information, which can better represent fine, subtle facial movements.

Considering deep learning methods have achieved good performances in facial expression recognition, recently, researchers have attempted to apply deep learning to the task of ME recognition. In [24], Kim et al. proposed to use convolutional neural network (CNN) to encode the spatial features of MEs at different expression-states, and then transfer the spatial features into a Long Short-Term Memory (LSTM) network to learn spatiotemporal features. Peng et al. [25] proposed a dual time-scale convolutional neural network, in which the different stream structures of the network can be used to adapt to ME clips of different frame rates. Li et al. [26] proposed spotting ME apex frames in the frequency domain and fine-tuning a VGG-Face model with magnified apex frames. In the work of [27], Khor et al. introduced an Enriched Long-term Recurrent Convolutional Network (ELRCN) model for micro-expression recognition, which encodes ME features by combining a deep spatial feature learning module and a temporal learning module. Li et al. [28] presented a 3D flow-based CNN (3D-FCNN) model for micro-expression recognition, which uses optical flow together with raw grayscale frames as input to a 12-layer deep network.

Due to the difficulties of ME elicitation and sample annotation, available datasets for training are very small, which limits the performances of deep neural networks for ME recognition. This article investigates the application of a shallow PCANet+ [29] model for the task of ME recognition. PCANet [30] combines principal component analysis (PCA) with CNN architecture. Despite its simplicity, PCANet has achieved promising results in image classification tasks, such as face recognition. As an extension model, PCANet+ eliminates the problem of complete linearity of PCANet and also alleviates the problem of feature dimension explosion by adding a pooling unit between adjacent layers. In this article, we propose a novel ME recognition method (OF-PCANet+) by incorporating the PCANet+ network and dense optical flow calculation. Considering the subtlety of MEs, we first calculate the optical flow from input ME video clips to enhance the motion information; then, we construct multi-channel images by stacking the optical flow fields of consecutive frames and feed them into a two-layer PCANet+ network to learn more powerful spatiotemporal features. A linear SVM is adopted in the classification of ME video clips. Experimental results on publicly available SMIC [12] and CASME2 [15] datasets demonstrate the effectiveness of the proposed method. The main contributions of this article are summarized as follows:

- We propose a lightweight OF-PCANet+ method for ME recognition, which is computationally simple and which can meanwhile produce promising recognition performance.
- We present a spatiotemporal feature learning strategy for ME recognition. Discriminative spatiotemporal features can be learned automatically by feeding stacked optical flow sequences into the PCANet+ network.

The rest of this article is organized as follows. Section 2 gives a brief introduction to optical flow calculation and the PCANet+ model. Section 3 describes our proposed method in detail. Section 4 presents experimental results and discussions, and the conclusions are given in Section 5.

2. Preliminaries

Table 1 shows the convention of variable representation adopted in this article. We express the sequential image data of MEs in two forms: (1) an intensity function $I : \mathbb{R}^3 \to \mathbb{R}$,

which takes three inputs corresponding to the spatial x, y components and the temporal t component, respectively; (2) a three-dimensional matrix $\mathbf{I} \in \mathbb{R}^{N \times M \times L}$, where N, M, L denote the height, width, and length of image data, respectively.

Table 1. Convention of variable representation.

Variable Symbol	Description
$\boldsymbol{a} \in \mathbb{R}^D$	A D-dimensional real vector.
$\boldsymbol{a}(i) \in \mathbb{R}$	i-th element of vector $\boldsymbol{a}$.
$\mathbf{A} \in \mathbb{R}^{N \times M}$	A 2-dimensional real matrix with N rows and M columns.
$\mathbf{A} \in \mathbb{R}^{N \times M \times L}$	A 3-dimensional real matrix with size of $N \times M \times L$.
$\mathbf{A}(i:j, k:l, n:m) \in \mathbb{R}^{(j-i+1)\times(l-k+1)\times(m-n+1)}$	A clipped matrix of $\mathbf{A} \in \mathbb{R}^{N \times M \times L}$, where $i, j, k, l, n, m \geq 1, i \leq j \leq N, k \leq l \leq M, n \leq m \leq L$.
$\mathcal{A}$	A set.
$\lvert\mathcal{A}\rvert \in \mathbb{N}$	Size of the set $\mathcal{A}$.

2.1. Optical Flow

Optical flow estimation methods take advantage of two assumptions: the constraint of brightness constancy and small motion. The brightness constancy assumes that the gray level of the moving object remains unchanged, and the small motion assumes that the velocity vector field changes very slowly in a short time interval. We suppose that a pixel $I(x, y, t)$ in a video clip will move by $\Delta x, \Delta y, \Delta t$ to the next frame. According to the constraint of brightness constancy mentioned above, the pixel intensity before and after movement is constant, and we can obtain

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t). \tag{1}$$

Based on the constraint of small motion. The right part of Equation (1) can be expanded by Taylor series, as below:

$$\begin{aligned} I(x + \Delta x, y + \Delta y, t + \Delta t) &= I(x, y, t) \\ &+ \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \varepsilon, \end{aligned} \tag{2}$$

where ε represents the high-order term, which can be ignored. Substitute it into Equation (1), we obtain:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0. \tag{3}$$

Let u and v represent the horizontal and vertical components of optical flow, respectively, as $u = \frac{\Delta x}{\Delta t}, v = \frac{\Delta y}{\Delta t}$. Substitute them into Equation (3), and we have

$$I_x u + I_y v + I_t = 0, \tag{4}$$

where $I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y}, I_t = \frac{\partial I}{\partial t}$ represent the partial derivatives of pixel intensity to x, y, and t, respectively, and (u, v) is called the optical flow field.

2.2. PCANet

For a gray-scale image input $\mathbf{I} \in \mathbb{R}^{N \times M}$, the PCANet extracts a $k_1 \times k_2$ patch around each pixel. Subtract each patch with its patch mean and then reshape it into a vector with length of $k_1 k_2$; we can obtain NM normalized patch vectors. By concatenating them to construct a matrix, we can obtain a normalized patch matrix of $\mathbf{I}$ as $\mathbf{P} \in \mathbb{R}^{k_1 k_2 \times NM}$, where each column denotes a single patch vector. Assume that we have a batch of B images; concatenating all patches generated from all of the images in the batch similarly gives the

patch matrix as $\mathbf{P} \in \mathbb{R}^{k_1 k_2 \times BNM}$. The PCANet aims to minimize a reconstruction error with respect to each patch, as follows.

$$\min_{\mathbf{V} \in \mathbb{R}^{k_1 k_2 \times L_1}} \|\mathbf{P} - \mathbf{V}\mathbf{V}^T\mathbf{P}\|_2^2, \text{s.t.} \mathbf{V}^T\mathbf{V} = \mathbb{I}_{L_1}, \tag{5}$$

where L_1 denotes the number of PCA filters and $\mathbb{I}_{L_1}$ denotes an identity matrix with size of $L_1 \times L_1$. This equation is actually a classic principal component analysis, whose solution is known as the L_1 principal eigenvectors of $\mathbf{P}\mathbf{P}^T$. Based on this, the l-th PCA filter is derived by reshaping the l-th principal eigenvectors of $\mathbf{P}\mathbf{P}^T$ into a $k_1 \times k_2$ matrix $\mathbf{W}_l$. For one PCANet layer with L_1 PCA filters, the output of the i-th image $\mathbf{I}_i \in \mathbb{R}^{N \times M}$ in the batch will be $\mathcal{O}_i = \left\{\mathbf{I}_i * \mathbf{W}_1, \mathbf{I}_i * \mathbf{W}_2, \ldots, \mathbf{I}_i * \mathbf{W}_{L_1}\right\}$, where $*$ denotes the convolution operation. Similarly, extracting patches from $\mathcal{O}_i$ and concatenating them like before, we obtain the input for the next layer $\mathbf{P}' \in \mathbb{R}^{k_1 k_2 \times L_1 BNM}$.

The PCANet could be constructed into a multi-layer architecture, but due to the problem of feature dimension explosion, it usually has many fewer layers than the normal deep neural networks. Here, we only consider a two-layer PCANet, which is widely used. It should be noted that before the final output, there will be a feature encoding layer with the application of hashing and histogram. Let $\mathbf{O}_k^1 = \mathbf{I}_i * \mathbf{W}_k^1 \in \mathbb{R}^{N \times M}$ be the output of the convolution operation in the 1st layer, where $\mathbf{W}_k^1$ denotes the k-th PCA filter in the 1st layer. Then, a hash map will be generated by the following equation to combine the output of each filter.

$$\mathbf{T}_l = \sum_{k=1}^{L_2} 2^{k-1} H(\mathbf{O}_l^1 * \mathbf{W}_k^2), \tag{6}$$

where L_2 denotes the number of PCA filters in the 2nd layer, $H(\cdot)$ is a Heaviside step function, whose value is one for positive entries and zero otherwise. $\mathbf{W}_k^2$ denotes the k-th PCA filter in the 2nd layer. Let $\text{Hist}(\cdot)$ be the function that outputs the histogram vector of the 2^{L_2} hash labels in a hash map. The final feature vector is expressed as

$$\boldsymbol{f}_i = \left[Hist(\mathbf{T}_1), Hist(\mathbf{T}_2), \ldots, Hist(\mathbf{T}_{L_1})\right]. \tag{7}$$

2.3. PCANet+

Because the PCANet layers are completely linearly connected, the lack of nonlinearity could decrease the feature learning effect. The PCANet+ overcomes this problem by adding a mean pooling layer between two consecutive layers, which also helps reduce the feature dimensions. The PCANet+ also extends the original network to support the input of multi-channel images.

Given a multi-channel image $\mathbf{I} \in \mathbb{R}^{N \times M \times F_{l-1}}$, where N, M denotes the height and the width, respectively. F_{l-1} denotes the number of channels of the input image, which could also denote the number of the filters of the previous layer. Similar to the PCANet, several three-dimensional patches with size of $k_l \times k_l \times F_{l-1}$ will be generated, where k_l denotes the filter size of the l-th layer. Thereafter, all of the patches will be reshaped as $\mathbf{P} \in \mathbb{R}^{k_l^2 F_{l-1} \times BNM}$, which is used for filter learning. Let F_l be the number of PCA filters of the current layer and let $\mathbf{W}_k^l \in \mathbb{R}^{k_l \times k_l \times F_{l-1}}$ be the k-th learned filter; the output of this layer is expressed as

$$\mathbf{I}' = \left[\beta(\mathbf{I} * \mathbf{W}_1^l), \beta(\mathbf{I} * \mathbf{W}_2^l), \ldots, \beta(\mathbf{I} * \mathbf{W}_{F_l}^l)\right] \in \mathbb{R}^{N \times M \times F_l}, \tag{8}$$

where $\beta(\cdot)$ denotes the mean pooling.

It should be noted that, for the feature encoding layer, based on the one in the PCANet, the PCANet+ also apply the chunking strategy on both the filter level and the image level.

For the computation of the hash map, the F_l outputs of the filters are divided into F_λ subsets; then, the hash map for each subset is computed as

$$\mathbf{T}_t^l = \sum_{f=1}^{F_\lambda} 2^{f-1} H(\beta(\mathbf{I} * \mathbf{W}_{(t-1)\times F_\lambda + f}^l)), \tag{9}$$

where $t = \{1, 2, \ldots, \frac{F_l}{F_\lambda}\}$ is the index of the subset. PCANet+ partitions each $\mathbf{T}_t^l$ into B_l nonoverlapping blocks, which is histogrammed into 2^{F_l} bins. Finally, the output of the feature encoding has a size of $\frac{F_l}{F_\lambda} B_l 2^{F_\lambda}$.

3. Method

In this section, we will describe the proposed method for micro-expression recognition in detail. Our method consists of three steps: (1) dense optical flow calculation and multi-channel stacking; (2) feature extraction with PCANet+; (3) classification with support vector machine. Figure 1 shows the overview of our proposed method.

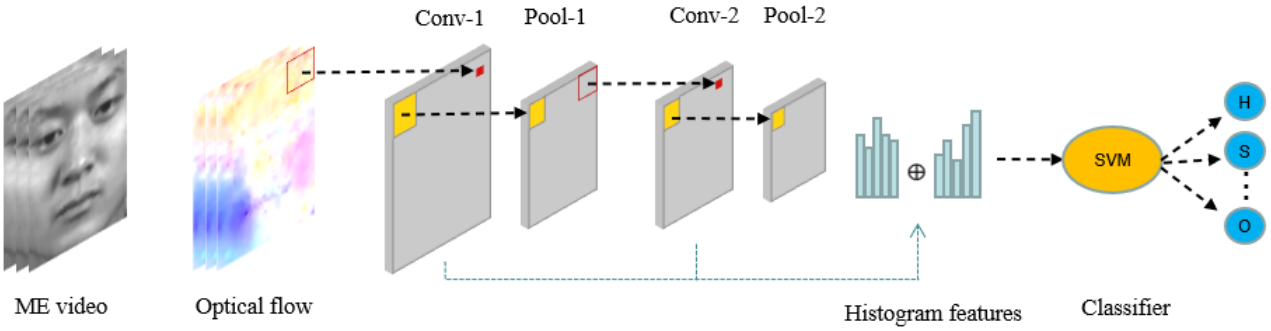

Figure 1. The framework of the proposed ME recognition method.

3.1. Dense Optical Flow Calculation and Multi-Channel Stacking

The optical flow is a two-dimensional vector field on image plane, which reflects the motion of pixels of two consecutive frames in a video sequence. In order to improve the effect of PCANet+ feature learning, we first perform a dense optical flow calculation on the original cropped ME video clips to enhance the facial motion information.

There are many methods for dense optical flow motion estimation. In this article, we apply the method presented in [31] to dense optical flow calculation, which introduces a subspace trajectory model to keep temporally consistent optical flow. For a single pixel of ME image data $I(x, y, t_0)$, to compute the sequential optical flow field $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{L-1}$ (L denotes the length of ME image sequence), they propose a loss function for optical flow estimation as follows.

$$\begin{aligned} E(x, y, t_0, \boldsymbol{u}, \boldsymbol{v}) = \alpha \iint_\Omega \textstyle\sum_{t=1}^{L} |I(x + \boldsymbol{u}(t), y + \boldsymbol{v}(t), t) - I(x, y, t_0)| dxdy \\ + \beta \iint_\Omega \textstyle\sum_{t=1}^{L} \| [\boldsymbol{u}(t), \boldsymbol{v}(t)] - \sum_{i=1}^{R} \boldsymbol{q}_i(t) \mathrm{lin}(\boldsymbol{u}(t), \boldsymbol{v}(t)) \|_2^2 dxdy \\ + \iint_\Omega \| \nabla \mathrm{lin}(\boldsymbol{u}(t), \boldsymbol{v}(t)) \|_2 dxdy, \end{aligned} \tag{10}$$

where $\boldsymbol{q}_1(t), \boldsymbol{q}_2(t), \ldots, \boldsymbol{q}_R(t) : \{1, 2, \ldots, L\} \to \mathbb{R}^2$ denote R basis trajectories used to construct the trajectory space. $\Omega \in \mathbb{R}^2$ denotes the image domain. $\mathrm{lin} : \mathbb{R}^2 \to \mathbb{R}^R$ denotes a map function that maps the optical flow field $\boldsymbol{u}(t), \boldsymbol{v}(t)$ to a new space constructed by the R basis trajectories. The first term is the penalty term of the brightness constancy constraint. The second term makes the derived optical flow lie on the basis trajectories. The third term is a total variation-based spatial regularization of the trajectory model coefficients.

Given an ME image sequence $\mathbf{I} \in \mathbb{R}^{N \times M \times L}$, we first set its first frame as the reference frame. Based on the optical flow motion estimation method above, we compute the optical flow field sequence of u and v components as $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{N \times M \times (L-1)}$. Figure 2 shows the

results of dense optical flow calculation for a ME video clip (happy class) of CASME2 dataset, in which Frame 1 is the reference frame, and we compute the optical flow field (UV1 to UV4) between the reference frame and the rest of the frames (Frame 2 to Frame 5) by a subspace trajectory model presented in [31]. It should be noted that we use color coding to illustrate the results of optical flow calculation. Different colors indicate different directions, and color saturation indicates the intensity of optical flow. It can be seen that optical flow field can better reflect the movement areas on the face, and it also has a certain effect on filtering the identity information of the face.

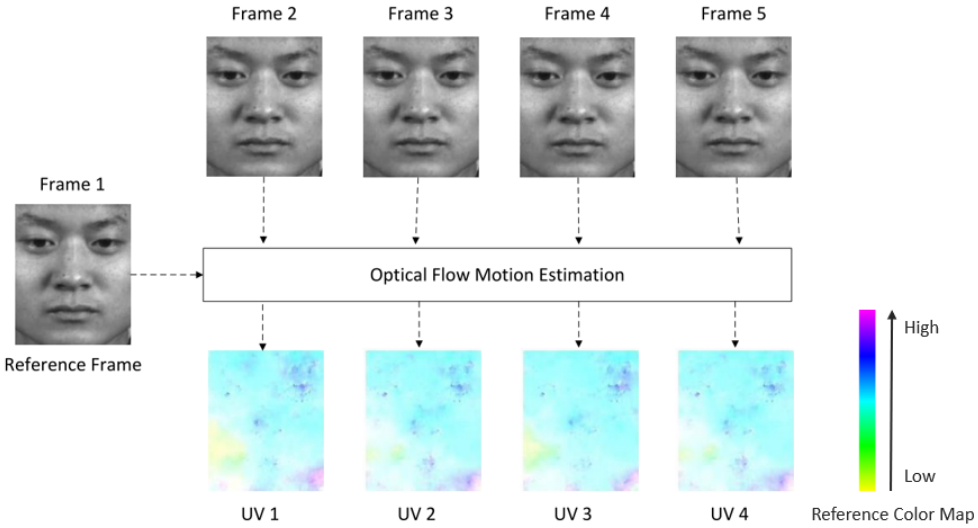

Figure 2. Example of optical flow motion estimation, where we set the first frame of ME image sequence as the reference frame and then compute the optical flow field between the reference frame and the rest of the frames with a subspace trajectory model.

To learn spatiotemporal features by PCANet+ based on optical flow, we conduct a multi-channel stacking operation on the optical flow sequences before they are fed to the PCANet+. Given the computed optical flow sequences $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{N \times M \times (L-1)}$, we use a sliding window with size of T and step size of s to sample them into several sequence subsets as

$$\begin{aligned} \mathcal{U} &= \left\{ \mathbf{U}_i \in \mathbb{R}^{N \times M \times T} : \mathbf{U}_i = \mathbf{U}(1:N, 1:M, (i-1)s+1 : T+(i-1)s), i \in [1, \lfloor \tfrac{L-T}{s} \rfloor] \right\} \\ \mathcal{V} &= \left\{ \mathbf{V}_i \in \mathbb{R}^{N \times M \times T} : \mathbf{V}_i = \mathbf{V}(1:N, 1:M, (i-1)s+1 : T+(i-1)s), i \in [1, \lfloor \tfrac{L-T}{s} \rfloor] \right\}, \end{aligned} \tag{11}$$

where $|\mathcal{U}| = |\mathcal{V}| = \lfloor \frac{L-T}{s} \rfloor$. Then, each element in $\mathcal{U}$ and $\mathcal{V}$ will be concatenated to form a stacked input sequence as

$$\mathcal{I} = \left\{ \mathbf{I}_i \in \mathbb{R}^{N \times M \times (2T)} : \mathbf{I}_i = \mathbf{U}_i \| \mathbf{V}_i, i \in [1, \lfloor \frac{L-T}{s} \rfloor], \mathbf{U}_i \in \mathcal{U}, \mathbf{V}_i \in \mathcal{V} \right\}, \tag{12}$$

where $\|$ denotes the matrix concatenating operation through the third dimension. Through the multi-channel stacking operation, the optical flow sequence for each video clip is converted into multi-channel images by stacking adjacent T frames in a sliding window, as shown in Figure 3. These multi-channel images will be fed to PCANet+ network to learn more discriminatory features.

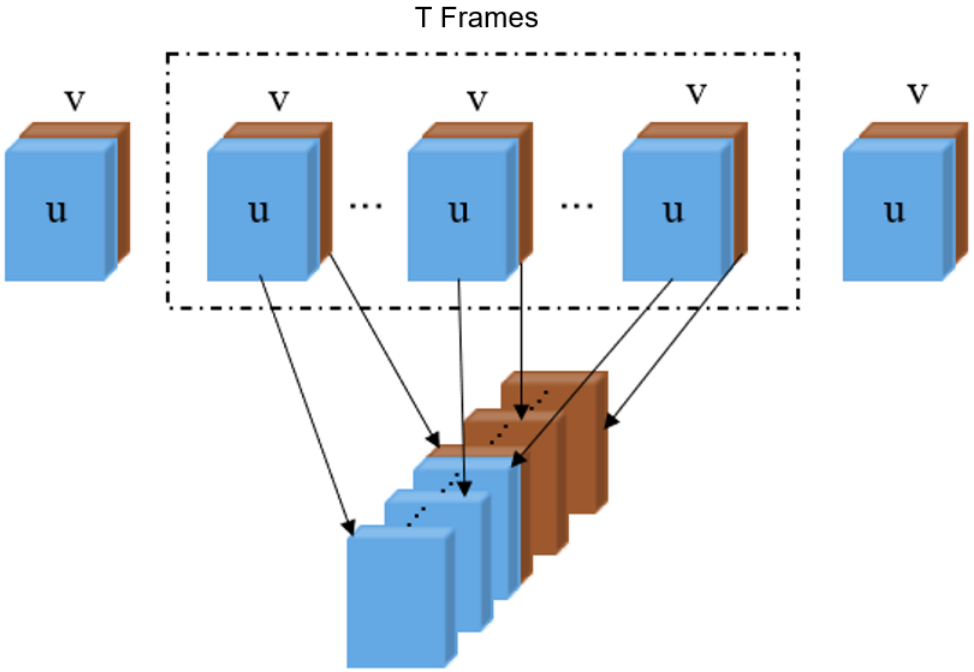

Figure 3. Illustration of stacking optical flow sequences into multi-channel images.

3.2. Feature Extraction with PCANet+

PCANet+ can take multi-channel images as input, which therefore makes the capacity of learned filter bank much larger than PCANet [29]. In this article, multi-channel images based on stacking of optical flow sequences are used as input to PCANet+ network for further feature extraction.

For K cropped video clips in dataset, after optical flow calculation and stacking process illustrated in Figure 3, we obtain a combined multi-channel image set $\mathcal{I}_{\text{all}} = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \ldots \cup \mathcal{I}_K$, where $\mathcal{I}_i$ denotes the multi-channel images of the i-th video clip. $|\mathcal{I}_{\text{all}}| = L_1 + L_2 + \ldots + L_K$, where L_i represents the number of multi-channel images generated from the ith video clip after stacking. Here, we set the step size of sliding window as $s = (T-1)/2$. Then, $\mathcal{I}_{\text{all}}$ will be fed to a 2-layer PCANet+ with D_1 filters (size: $k_1 \times k_1$) in the 1st layer and D_2 filters (size: $k_2 \times k_2$) in the 2nd layer. To facilitate the succeeding binary hash coding stage in PCANet+, the number of filters D_1, D_2 need to be configured to a multiple of D_λ. According to [29], we prefix $D_\lambda = 8$ in our experiments. Slightly different from the original PCANet+, we apply feature encoding to each PCANet+ layer and concatenate their outputs as the final feature representation, which has $\sum_{l=1}^{2} B_l \frac{F_l}{F_\lambda} 2^{F_\lambda}$ dimensions in total. Finally, a linear SVM is adopted in the classification of ME video clips.

4. Experimental Results and Analysis

To evaluate the proposed method for micro-expression recognition, we conduct comprehensive experiments on two widely used ME datasets, SMIC and CASME2. We first introduce the datasets and evaluation metrics used in experiments, and then we present the experimental results and discussions.

4.1. Settings

The SMIC [12] provides three data subsets with different types of recording cameras: SMIC-HS, SMIC-VIS, and SMIC-NIR. SMIC-VIS and SMIC-NIR were recorded by normal speed cameras with 25 fps of visual (VIS) and near infrared (NIR) light range, respectively. Because MEs are rapid facial movements, high speed cameras help to capture more temporal information. In our experiments, the SMIC-HS subset recorded by 100 fps high-speed cameras is used, which contains 164 spontaneous facial ME video clips from

16 subjects. These samples are divided into three ME classes: positive (51 samples), negative (70 samples), and surprise (43 samples).

The CASME2 [15] dataset consists of 247 spontaneous facial ME video clips with spatial resolution 640 × 480. This dataset was collected by a high-speed camera at 200 fps. As well, MEs of participants were elicited in a well-controlled laboratory environment with four lamps providing steady and high-intensity illumination. The CASME2 dataset includes five ME classes: happiness (32 samples), surprise (25 samples), disgust (64 samples), repression (27 samples), and others (99 samples). The frames of a sample video clip (happiness) in the CASME2 dataset are shown in Figure 4.

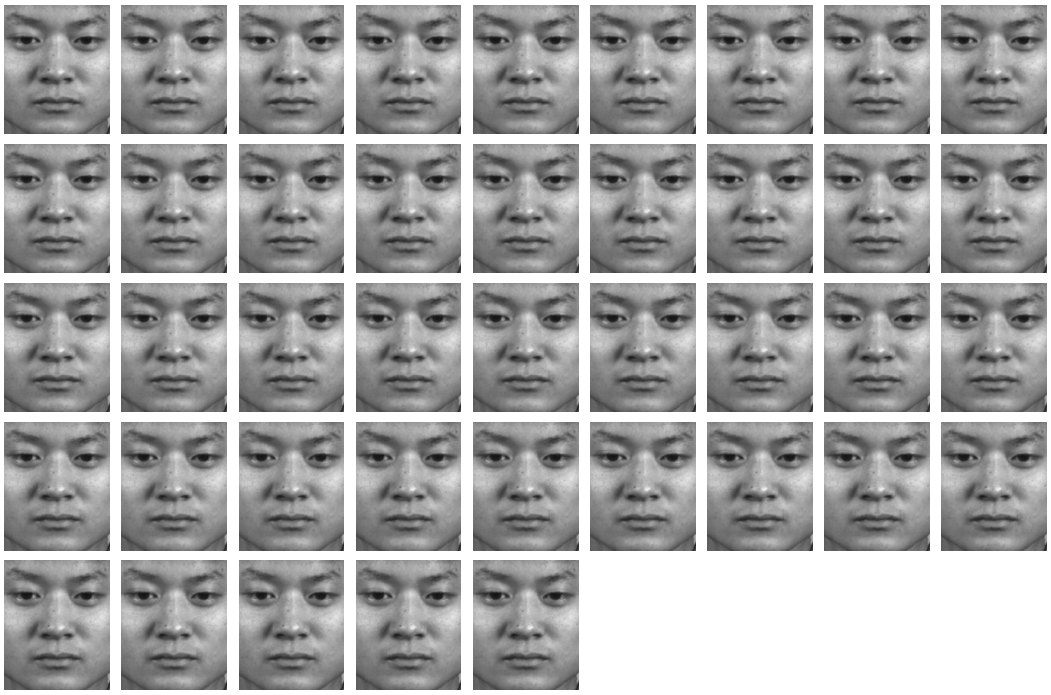

Figure 4. The frames of a sample video clip (happiness) in CASME2 dataset.

The characteristics of two public datasets used in our experiments are summarized in Table 2. To set up a person-independent configuration, leave-one-subject-out (LOSO) cross validation protocol is adopted, where the samples from one subject are used as the testing set, and the samples from the remaining subjects are used as the training set. A linear SVM based on features extracted from PCANet+ is adopted in the classification stage.

Table 2. Detailed information of SMIC and CASME2 dataset.

Dataset	SMIC-HS	CASME2
Subjects	16	26
Sample	164	247
Year	2013	2014
Frame Rate	100	200
Image Resolution	640 × 480	640 × 480
Emotion classes	3 categories: positive (51) negative (70) surprise (43)	5 categories: happiness (32) surprise (25) disgust (64) repression (27) others (99)

Performance metrics such as accuracy, Macro-F1, and Macro-recall, are used in evaluation. Macro-F1 and Macro-recall represent the average F1-score and recall of all classes.

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} FP_i} \tag{13}$$

$$\text{P}_i = \frac{TP_i}{TP_i + FP_i} \tag{14}$$

$$\text{R}_i = \frac{TP_i}{TP_i + FN_i} \tag{15}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times \text{P}_i \times \text{R}_i}{\text{P}_i + \text{R}_i} \tag{16}$$

$$\text{Macro-recall} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{17}$$

where C is the class number and TP_i, FP_i, and FN_i represent true positive, false positive, and false negative of class i, respectively.

4.2. Effects of Parameters in PCANet+

We need to investigate the hyper-parameters in the OF-PCANet+ method, including the number of frames in stacking (T) and the size and number of filters ($[k_1, D_1][k_2, D_2]$). In this article, we build a two-layer PCANet+ model in our method, based on the observation that deeper architectures will not necessarily lead to further performance improvements. In this section, we conduct experiments to examine the influence of these parameters on recognition performance.

4.2.1. The Number of Frames in Stacking

We first examine the number of frames (T) in the process of stacking optical flow sequences. In this experiment, the filter size and number of the network are set to $[k_1, D_1] = [7, 32]$, $[k_2, D_2] = [9, 16]$. Table 3 reports the effect of frame stacking number T on recognition accuracy.

Table 3. ME recognition results of OF-PCANet+ with respect to different frame stacking number, T.

Frame Stacking Number T	SMIC			CASME2		
	Accuracy	**Macro-F1**	**Macro-Recall**	**Accuracy**	**Macro-F1**	**Macro-Recall**
1	0.4268	0.3924	0.3890	0.2301	0.2437	0.2316
3	0.6159	0.6184	0.6214	0.4959	0.4960	0.4786
5	0.6280	0.6309	0.6369	0.5203	0.5266	0.5148
7	0.6098	0.6109	0.6131	0.4512	0.4412	0.4270

As shown in Table 3, the performances can be improved by using the operation of frame stacking compared with non-stacking ($T = 1$). The results indicate that multi-frame stacking of optical flow sequences can help the PCANet+ network learn spatiotemporal information, which is very important in ME recognition. When stacking number T increases from 1 (i.e., no stacking) to 5, the performances become better, and when T increases to 7, the recognition accuracies start to decrease. In the following experiments, we set the best frame stacking number as $T = 5$.

4.2.2. The Size and Number of Filters in Each Layer

We next do experiments to examine the number and size of filters $[k, D]$ used in the OF-PCANet+. The performances in terms of accuracy, macro-F1, and macro-recall with different combinations of $[k, D]$ are reported in Table 4, where $k \in \{5, 7, 9, 11, 13, 15\}$ and

$D \in \{8, 16, 32\}$. We can see that the proposed method achieves the best recognition performances (in bold) under settings of $[k_1, D_1] = [7, 32]$, $[k_2, D_2] = [9, 16]$ on the SMIC dataset and $[k_1, D_1] = [7, 16]$, $[k_2, D_2] = [7, 32]$ on the CASME2 dataset. In Table 5, we summarize the best configuration of the PCANet+ network in our method. Figure 5 presents the visualization of feature maps with the parameter of $[k_1, D_1] = [7, 16]$, $[k_2, D_2] = [9, 16]$ produced in layer 1 and layer 2, respectively, for an input video clip from the CASME2 dataset. The bright areas have higher motion energy, which means that the facial movements are relatively strong around these areas.

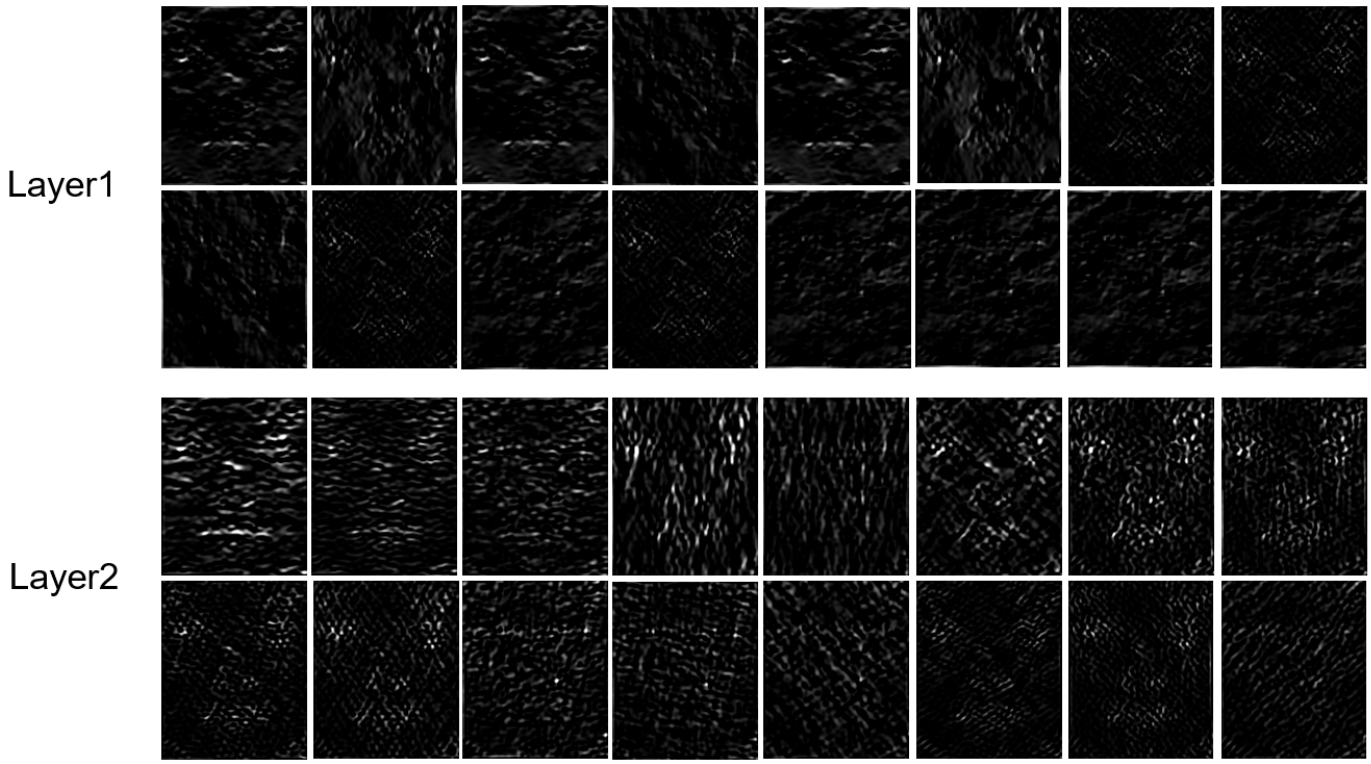

Figure 5. The visualization of feature maps produced in each layer for an input video clip from CASME2 dataset.

Table 4. ME recognition results of OF-PCANet+ with respect to different number and size of filters $[k, D]$.

$[k_1, D_1][k_2, D_2]$	SMIC			CASME2		
	Accuracy	**Macro-F1**	**Macro-Recall**	**Accuracy**	**Macro-F1**	**Macro-Recall**
[5, 16][7, 16]	0.5854	0.5893	0.5941	0.5000	0.5122	0.4962
[5, 32][5, 8]	0.5854	0.5880	0.5941	0.5041	0.5047	0.4950
[5, 32][5, 16]	0.5915	0.5954	0.6036	0.5081	0.5114	0.5020
[5, 32][5, 32]	0.5793	0.5834	0.5905	0.5163	0.5198	0.5055
[7, 16][9, 32]	0.6098	0.6127	0.6173	0.5285	0.5272	0.5031
[7, 32][5, 16]	0.5976	0.6010	0.6084	0.5122	0.5128	0.4950
[7, 32][7, 16]	0.6098	0.6137	0.6209	0.5041	0.5081	0.4867
[7, 32][9, 16]	**0.6280**	**0.6309**	**0.6369**	0.5203	0.5266	0.5148
[7, 16][7, 32]	0.6037	0.6046	0.6096	**0.5325**	**0.5493**	**0.5241**
[7, 32][11, 16]	0.6037	0.6053	0.6126	0.5285	0.5280	0.5067
[7, 32][13, 16]	0.5976	0.6007	0.6048	0.4919	0.4931	0.4724
[7, 32][15, 16]	0.6220	0.6247	0.6310	0.5081	0.5152	0.4931
[9, 16][11, 16]	0.5915	0.5943	0.6001	0.4268	0.4096	0.4096
[13, 16][15, 16]	0.6098	0.6131	0.6167	0.4350	0.4250	0.4250

Table 5. Summary of the configuration of PCANet+ network.

	Best Configuration for SMIC	**Best Configuration For CASME2**
$\boldsymbol{I}_i$	$170 \times 139 \times 10$	$170 \times 139 \times 10$
$\boldsymbol{W}_1$ $(k_1 \times k_1 \times 2T) \times D_1$	$(7 \times 7 \times 10) \times 32$ Str. 1, Pad. 3	$(7 \times 7 \times 10) \times 16$ Str. 1, Pad. 3
Pool-1	3×3 Mean Pooling, Str. 1	3×3 Mean Pooling, Str. 1
$\boldsymbol{W}_2$ $(k_2 \times k_2 \times 2T) \times D_2$	$(9 \times 9 \times 32) \times 16$ Str. 1, Pad. 4	$(7 \times 7 \times 16) \times 32$ Str. 1, Pad. 3

4.3. Comparison with Other Methods

To demonstrate the effectiveness of OF-PCANet+, we compare the method with some existing handcrafted methods as well as deep learning methods. The size and number of filters in layer 1 and layer 2 are set to $[k_1, D_1] = [7, 32]$, $[k_2, D_2] = [9, 16]$ for SMIC and $[k_1, D_1] = [7, 16]$, $[k_2, D_2] = [7, 32]$ for CASME2. Following the experiment settings of [12,15], we re-implement LBP-TOP with 8×8 and 5×5 facial blocks, radius $[R_{XY}, R_{XT}, R_{YT}] = [4, 1, 1]$. For STLBP-IP, the block size of 4×7 is used for the SMIC dataset, and 8×9 for the CASME2 dataset, as suggested in [18].

Table 6 reports the results of performance comparison of different methods in terms of accuracy, macro-F1, and macro-recall on the SMIC and CASME2 datasets, where N/A indicates that the corresponding performance was not given in the article. We can see that the proposed OF-PCANet+ model outperforms popular hand-crafted methods, i.e., LBP-TOP, STLBP-IP, and KGSL, both on SMIC and CASME2. Furthermore, our method also shows comparable performances with deep learning methods, such as ELRCN [27] and 3D-FCNN [28]. The results indicate that the shallow model of PCANet+ can learn effective spatiotemporal features of micro-expressions based on multi-frame stacking of optical flow sequences.

Table 6. Comparisons of different methods.

Method	**SMIC**			**CASME2**		
	Accuracy	**Macro-F1**	**Macro-Recall**	**Accuracy**	**Macro-F1**	**Macro-Recall**
LBP-TOP [15]	0.4207	0.4266	0.4429	0.4390	0.4297	0.4259
STLBP-IP [18]	0.4329	0.4270	0.4241	0.4173	0.4026	0.4282
KGSL [19]	0.5244	0.4937	0.5162	0.4575	0.4325	0.4437
ELRCN [27]	N/A	N/A	N/A	0.5244	0.5000	0.4396
3D-FCNN [28]	0.5549	N/A	N/A	0.5911	N/A	N/A
OF-PCANet+	0.6280	0.6309	0.6369	0.5325	0.5493	0.5241

5. Conclusions

In this article, we propose a simple yet effective method OF-PCANet+ for micro-expression recognition by incorporating the dense optical flow calculation with a shallow PCANet+ network. By multi-frame stacking of optical flow sequences as input, discriminative spatiotemporal features can be learned by a two-layer PCANet+ model. Moreover, because the filters can be learned analytically only with the PCA algorithm in each layer, the training process of our method is much simpler than deep learning methods based on back propagation algorithm. The experimental results on SMIC and CASME2 datasets demonstrate the promising performance of the proposed method. In future work, we will try to apply this method to other related tasks, such as behavior recognition and video classification.

Author Contributions: Conceptualization, S.W. and F.L.; methodology, S.G. and H.L.; software, S.G. and H.L.; validation, J.H. and J.Y.; formal analysis, F.L. and J.H.; investigation, S.W.; resources, F.L. and J.Y.; data curation, H.L.; writing—original draft preparation, S.W.; writing—review and editing, J.H. and F.L.; visualization, S.G.; supervision, J.Y. and F.L.; project administration, F.L.; funding acquisition, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Science Foundation of Fujian Province of China (No. 2019J01001) and the Industry-University-Research Project of Xiamen City (3502Z20203002).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

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

References

1. Bhushan, B. Study of Facial Micro-expressions in Psychology. In *Understanding Facial Expressions in Communication: Cross-Cultural and Multidisciplinary Perspectives*; Springer: New Delhi, India, 2015; pp. 265–286.
2. Shen, X.; Wu, Q.; Fu, X. Effects of the duration of expressions on the recognition of microexpressions. *J. Zhejiang Univ. Sci. B* **2012**, *13*, 221–230. [CrossRef] [PubMed]
3. Yan, W.J.; Wu, Q.; Liang, J.; Chen, Y.H.; Fu, X. How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions. *J. Nonverbal Behav.* **2013**, *37*, 217–230. [CrossRef]
4. Xu, F.; Zhang, J. Facial Microexpression Recognition: A Survey. *Acta Autom. Sin.* **2017**, *43*, 333–348.
5. Ekman, P.; Friesen, W.V. Nonverbal Leakage and Clues to Deception. *Psychiatry* **1969**, *32*, 88–106. [CrossRef] [PubMed]
6. Ekman, P. Lie Catching and Micro Expressions. In *The Philosophy of Deception*; Martin, C., Ed.; Oxford University Press: Oxford, UK, 2009; pp. 118–133.
7. O'Sullivan, M.; Frank, M.; Hurley, C.; Tiwana, J. Police Lie Detection Accuracy: The Effect of Lie Scenario. *Law Hum. Behav.* **2009**, *33*, 530. [CrossRef] [PubMed]
8. Corneanu, C.A.; Simón, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [CrossRef] [PubMed]
9. Li, Y.; Huang, X.; Zhao, G. Joint Local and Global Information Learning with Single Apex Frame Detection for Micro-Expression Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 249–263. [CrossRef] [PubMed]
10. Warren, G.; Schertler, E.; Bull, P. Detecting Deception from Emotional and Unemotional Cues. *J. Nonverbal Behav.* **2009**, *33*, 59–69. [CrossRef]
11. Yan, W.J.; Wang, S.J.; Liu, Y.J.; Wu, Q.; Fu, X. For micro-expression recognition: Database and suggestions. *Neurocomputing* **2014**, *136*, 82–87. [CrossRef]
12. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–6.
13. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognising spontaneous facial micro-expressions. In Proceedings of the International Conference on Computer Vision, Washington, DC, USA, 6–13 November 2011; pp. 1449–1456.
14. Zhao, G.; Pietikainen, M. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef] [PubMed]
15. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. *PLoS ONE* **2014**, *9*, e86041. [CrossRef] [PubMed]
16. Huang, X.; Zhao, G.; Hong, X.; Pietikäinen, M.; Zheng, W. Texture Description with Completed Local Quantized Patterns. In Proceedings of the Scandinavian Conference on Image Analysis, Espoo, Finland, 17–20 June 2013; pp. 1–10.
17. Wang, Y.; See, J.; Phan, R.C.W.; Oh, Y.H. LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition. In Proceedings of the Asia Conference on Computer Vision 2014, Singapore, 1–5 November 2014; Springer International Publishing: Cham, Switzerland, 2015; pp. 525–537.
18. Huang, X.; Wang, S.; Zhao, G.; Piteikainen, M. Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 1–9.
19. Zong, Y.; Huang, X.; Zheng, W.; Cui, Z.; Zhao, G. Learning From Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition. *IEEE Trans. Multimed.* **2018**, *20*, 3160–3172. [CrossRef]
20. Lu, Z.; Luo, Z.; Zheng, H.; Chen, J.; Li, W. A Delaunay-Based Temporal Coding Model for Micro-expression Recognition. In Proceedings of the Asia Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 698–711.

21. Xu, F.; Zhang, J.; Wang, J.Z. Microexpression Identification and Categorization Using a Facial Dynamics Map. *IEEE Trans. Affect. Comput.* **2017**, *8*, 254–267. [CrossRef]
22. 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. *IEEE Trans. Affect. Comput.* **2016**, *7*, 299–310. [CrossRef]
23. Liong, S.T.; See, J.; Phan, R.C.W.; Oh, Y.H.; Cat Le Ngo, A.; Wong, K.; Tan, S.W. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Process. Image Commun.* **2016**, *47*, 170–182. [CrossRef]
24. Kim, D.H.; Baddar, W.J.; Ro, Y.M. Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations. In Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 15–19 October 2016; pp. 382–386.
25. Peng, M.; Wang, C.; Chen, T.; Liu, G.; Fu, X. Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition. *Front. Psychol.* **2017**, *8*, 1745. [CrossRef] [PubMed]
26. Li, Y.; Huang, X.; Zhao, G. Can Micro-Expression be Recognized Based on Single Apex Frame? In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3094–3098.
27. 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 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 667–674.
28. Li, J.; Wang, Y.; See, J.; Liu, W. Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Anal. Appl.* **2019**, *22*, 1331–1339. [CrossRef]
29. Low, C.Y.; Teoh, A.B.J.; Toh, K.A. Stacking PCANet+: An Overly Simplified ConvNets Baseline for Face Recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 1581–1585. [CrossRef]
30. Chan, T.H.; Jia, K.; Gao, S.; Lu, J.; Zeng, Z.; Ma, Y. PCANet: A Simple Deep Learning Baseline for Image Classification. *IEEE Trans. Image Process.* **2015**, *24*, 5017–5032. [CrossRef] [PubMed]
31. Garg, R.; Roussos, A.; Agapito, L. Robust Trajectory-Space TV-L1 Optical Flow for Non-rigid Sequences. In Proceedings of the Energy Minimization Methods in Computer Vision and Pattern Recognition, Petersburg, Russia, 25–27 July 2011; pp. 300–314.