Computational Visual Media

Volume 7 | Issue 3 Article 9

2021

Multi-scale joint feature network for micro-expression recognition

Xinyu Li
School of Software, Shandong University, Jinan 250101, China

Guangshun Wei
School of Software, Shandong University, Jinan 250101, China

Jie Wang
School of Software, Shandong University, Jinan 250101, China

Yuanfeng Zhou
School of Software, Shandong University, Jinan 250101, China

Computational Visual Media
https://doi.org/10.1007/s41095-021-0217-9 Vol. 7, No. 3, September 2021, 407–417
Research Article

Multi-scale joint feature network for micro-expression recognition

Xinyu Li[1], Guangshun Wei[1], Jie Wang[1], and Yuanfeng Zhou[1] (✉)

Abstract Micro-expression recognition is a substantive cross-study of psychology and computer science, and it has a wide range of applications (e.g., psychological and clinical diagnosis, emotional analysis, criminal investigation, etc.). However, the subtle and diverse changes in facial muscles make it difficult for existing methods to extract effective features, which limits the improvement of micro-expression recognition accuracy. Therefore, we propose a multi-scale joint feature network based on optical flow images for micro-expression recognition. First, we generate an optical flow image that reflects subtle facial motion information. The optical flow image is then fed into the multi-scale joint network for feature extraction and classification. The proposed joint feature module (JFM) integrates features from different layers, which is beneficial for the capture of micro-expression features with different amplitudes. To improve the recognition ability of the model, we also adopt a strategy for fusing the feature prediction results of the three JFMs with the backbone network. Our experimental results show that our method is superior to state-of-the-art methods on three benchmark datasets (SMIC, CASME II, and SAMM) and a combined dataset (3DB).

Keywords micro-expression recognition; multi-scale feature; optical flow; deep learning

1 Introduction

Micro-expressions are brief facial expressions that people unconsciously make when they try to hide a real emotion. They usually appear when people are in a critical situation. Initially, micro-expressions were defined as instant facial expressions during psychotherapy or micro-momentary facial expressions in nonverbal communication [1]. Later, Ekman and Friesen [2] analyzed a conversation video between a psychiatrist and a depressed patient, and observed that painful expressions occasionally appeared during the patient's smile, which he called micro-expressions. This was the first time that the term micro-expression was used. At the same time, related researchers also defined micro-expressions as facial expressions that can spontaneously reveal the real emotions of people, and are difficult to disguise. Therefore, micro-expression recognition has a wide range of applications in psychological and clinical diagnosis, emotional analysis, criminal investigation, and national defense security.

Micro-expression recognition is difficult, because (i) the duration of micro-expressions is very short, under 1/5 second, (ii) they only appear in specific areas of the face, and the intensity of the change is very weak, and (iii) existing micro-expression samples are scarce and numbers of samples in different classes are not balanced. Therefore, fully capturing the features of micro-expressions and improving recognition accuracy are still very challenging problems.

Since the task of micro-expression recognition was proposed, many associated methods have been suggested. In earlier years, Ekman [3] developed a micro-expression training tool (METT) to improve people's perception of micro-expressions. However, even with the help of METT, the recognition accuracy of professionals is still very low. With the development of computer vision and image processing technology, some researchers have used feature descriptors to extract features from images [4–6]. However, it is difficult to accurately capture tiny facial motion information by directly extracting features from the original images. Therefore,

1 School of Software, Shandong University, Jinan 250101, China. E-mail: X. Li, xinyuli@mail.sdu.edu.cn; G. Wei, guangshunwei@gmail.com; J. Wang, chiehwang@mail.sdu.edu.cn; Y. Zhou, yfzhou@sdu.edu.cn (✉).

Manuscript received: 2021-01-25; accepted: 2021-02-25

the work of Refs. [7–9] extracts micro-expression features by calculating facial optical flow. In addition, early deep learning models combined convolutional neural networks (CNN) and long short-term memory (LSTM) for micro-expression recognition. The complexity of this model can easily lead to overfitting problems in the training process when there are insufficient samples. Transfer learning and data augmentation techniques were introduced to overcome the difficulty of insufficient training samples [10, 11], but this undoubtedly increases the cost of model training. At present, some research works [12, 13] tend to use shallow multi-stream networks to improve the performance of the model on small datasets and in class-imbalanced situations.

In this paper, we ignore brute force solution methods that use different serial network combinations or multi-stream structures to improve the performance of model recognition. Instead, we use features in different layers in a single convolutional neural network, and propose a multi-scale joint feature network based on optical flow images. Since the intensity of micro-expressions in various samples is different, effective features of samples with weak intensity can be obtained in low layers but may be lost in high layers. However, samples with relatively high intensity can extract more effective features in the high layers of the network. Therefore, we hypothesize that features at different layers contribute to the classification of the network. To make better use of features from different layers, we propose a joint feature module. Further, we design two fusion strategies and observe the impact of different strategies on classification performance through experiments. Finally, we compare the proposed method with state-of-the-art methods; the results show that the performance of our model is competitive on three benchmark datasets and a cross-dataset.

2 Related work

In early stages, the slow development of micro-expression recognition was mainly due to the lack of a well-established database. The samples in an early micro-expression recognition database [5, 14] were simulated, non-spontaneous expressions of subjects. With the establishment of spontaneous micro-expression databases, SMIC [15], CASME [16], CASME II [17], CAS(ME)^2, and SAMM [18], relevant studies increased. At present, existing methods are mainly divided into two categories based on traditional methods and deep learning methods.

2.1 Traditional methods

Traditional methods generally combine handcrafted feature extraction with classical machine learning methods. Zhao and Pietikainen [6] used local binary patterns on three orthogonal planes (LBP-TOP) to enhance time dimension features. Subsequently, to enrich features in the time domain, completed local quantized patterns (CLQP) [19] and local binary patterns with six intersection point (LBP-SIP) [20] methods were successively proposed. Ben et al. [21] proposed three different methods using binary face descriptors, in which DCP-TOP based on dual-cross patterns and HWP-TOP based on hot wheel patterns encode features from micro-expression sequences. Wang et al. [22] considered the influence of color space on feature extraction and proposed a tensor independent color space (TICS) method to improve the performance of micro-expression recognition. Huang et al. [23] presented a method based on spatiotemporal local binary patterns and improved integral projection to extract facial feature information and distinguish feature information between different micro-expression classes.

To capture subtle facial motion information, optical flow [24, 25] was introduced into the field of micro-expression recognition. Liu et al. [8] put forward the main directional mean optical-flow (MDMO) method based on regions of interest, which reduced the dimensionality of features and improved the robustness of micro-expression recognition. As MDMO can cause loss of the underlying manifold structure, Liu et al. [26] proposed sparse MDMO, which constructs an effective sparse representation of micro-expressions by incorporating a new metric into GraphSC. Xu et al. [9] proposed facial dynamic maps (FDM), which took the optical flow field as the basis for extracting features of micro-expressions at different granularity. Liong et al. [7] came up with a new feature extractor, bi-weighted oriented optical flow (Bi-WOOF), which weighted the optical flow direction histogram twice to highlight facial motion, while proving that the micro-expression apex frame can provide sufficiently meaningful feature expressions.

2.2 Deep learning methods

Deep learning can directly learn hierarchical visual features from images, which has a wide range of applications to image classification and recognition, such as face recognition [27] and facial expression recognition [28]. Direct feature extraction from an original micro-expression sequence was attempted early on. Kim et al. [29] adopted a CNN to encode spatial features, and then used LSTM to extract temporal features of micro-expressions from continuous spatial features. However, the serial combination of these two deep networks undoubtedly increases the complexity of the model, with a tendency to overfitting for small datasets which are typical for micro-expressions. To avoid the problem of insufficient data, Peng et al. [30] and Wang et al. [31] both used a macro-expression dataset to train the network to obtain a preprocessing model and then applied it to the micro-expression recognition task. In addition to a macro-expression dataset for pre-training, Wang et al. [32] also used 560 micro-expression video clips from three micro-expression datasets to expand the network training data. Xia et al. [11] expanded micro-expression data based on multi-scale data augmentation by Eulerian video magnification (EVM) [33]. Quang et al. [34] used both transfer learning and data augmentation techniques to reduce the risk of overfitting during model training.

Due to the short duration of micro-expressions and their weak intensity, using the original image as input is not an effective way for the network to extract useful features. Therefore, to emphasise the main features, as a vector describing the motion information of objects, optical flow is widely used in deep learning methods for micro-expression recognition. In such works [10, 35], a CNN and LSTM are combined to extract micro-expression features from the original image and the optical flow image generated by the micro-expression motion. Inspired by Ref. [7], some subsequent studies only use the onset frame and the apex frame to represent a micro-expression sequence to reduce the computational cost of the network by avoiding redundant data. Liu et al. [36] also adopted adversarial training and EVM based on a pre-trained Resnet18 to improve the accuracy of micro-expression recognition. In addition, recent studies tend to design shallow networks and improve the performance of the model by increasing the number of branches of the network. For example, Refs. [12, 13, 37] designed a dual-stream network for micro-expression recognition, and Liong et al. [38] proposed a shallow triple stream three-dimensional CNN (STSTNet) for feature extraction and classification.

3 Method

To establish effective micro-expression recognition features, a novel micro-expression recognition framework is designed in this paper (see Fig. 1). In our micro-expression sequence, we first generate an RGB optical flow image (Section 3.1) that can describe the facial changes in a micro-expression, and then feed the optical flow image to the multi-scale joint feature network (Section 3.2) for feature extraction and classification.

3.1 Optical flow image

The micro-expression samples in the datasets were collected by high-speed cameras (100–200 fps), and each micro-expression sequence can be decomposed into multiple frames. The short duration and subtle intensity of micro-expressions make the changes between consecutive frames in the video sequence indistinct, that is, each frame is very similar. If we input all frames of a video sequence directly into the network, it will not only increase the computational cost but also result in feature redundancy. Instead, inspired by Refs. [7, 39], we only select the onset frame and the apex frame to represent each micro-expression sequence. Relative to the onset frame, the apex frame has the most obvious micro-expression intensity, but the change is extremely subtle compared to those of macro-expressions, and the number of samples in the micro-expression dataset is low. If the onset frame and apex frame are input to the network directly, it will be difficult for the network to learn effective features. Therefore, we use the onset frame and the apex frame to derive an optical flow image representing the dynamic changes in the micro-expression as the input of the network.

Optical flow, as a two-dimensional vector that describes changes in pixels between two images over time, can capture subtle changes in the face. The calculation of optical flow requires two basic assumptions: brightness constancy, and continuous motion or small motion. For a micro-expression

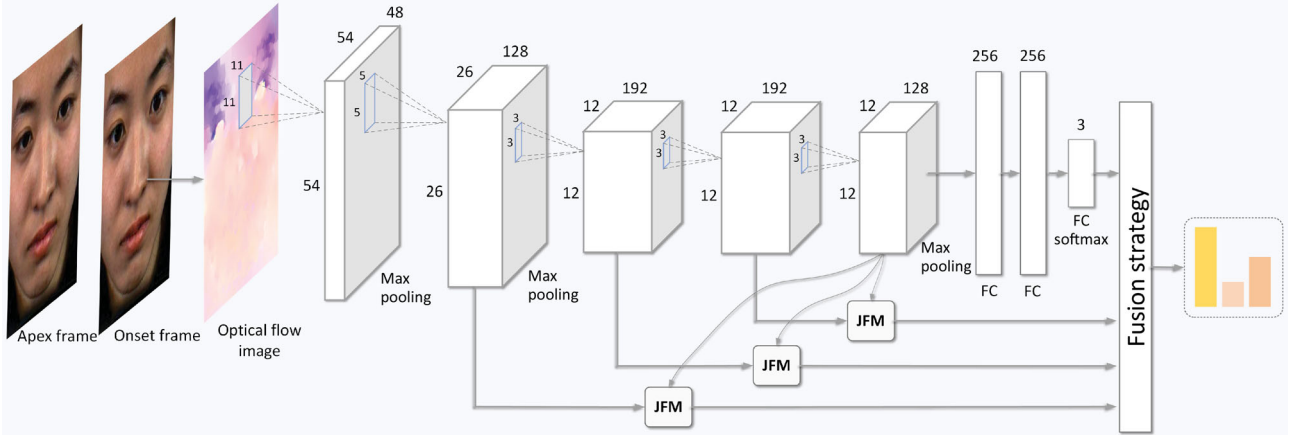

Fig. 1 Framework of our multi-scale joint network. Given a micro-expression video sequence, we employ the onset frame and the apex frame to obtain an optical flow image and feed it into the multi-scale joint feature network for feature extraction and classification.

sequence, the gray value of pixel k at location (x, y) in the onset frame is $G(x, y, t)$. Using the first assumption, we can obtain the gray value of pixel k in the apex frame:

$$G(x + \delta x, y + \delta y, t + \delta t) = G(x, y, t) \tag{1}$$

where δx and δy represent the distance pixel k has moved in the horizontal and vertical directions after a time δt.

Using the second assumption, the left hand side of Eq. (1) can be expanded as a first-order Taylor series, giving:

$$G(x, y, t) + \frac{\partial G}{\partial x}\delta x + \frac{\partial G}{\partial y}\delta y + \frac{\partial G}{\partial t}\delta t + \epsilon = G(x, y, t) \tag{2}$$

where ϵ represents a negligible higher order term. Therefore:

$$\frac{\partial G}{\partial x}u + \frac{\partial G}{\partial y}v + \frac{\partial G}{\partial t} = 0 \tag{3}$$

where $u = \delta x/\delta t$, $v = \delta y/\delta t$ represent the horizontal and vertical components of optical flow, respectively. The optical flow $\boldsymbol{u}_k$ can then be expressed as

$$\boldsymbol{u}_k = [u, v]^{\mathrm{T}} \tag{4}$$

Further, we can obtain for pixel k its optical flow magnitude $m_k = \sqrt{u^2 + v^2}$ and normalized direction $\theta_k = \arctan(v/u)/\pi$. In our work, we use TV-L1 [25] to calculate the optical flow between the onset frame and the apex frame of each sample. Then the pixel value at location (x, y) for channel $c \in \{R, G, B\}$ in the optical flow image can be written as

$$I_{\mathrm{o}}(x,\ y,\ c) = C(\widetilde{\theta}_k,\ c,\ \widetilde{m}_k) \tag{5}$$

where

$$\widetilde{\theta}_k = \frac{\theta_k + 1}{2}(n_c - 1) + 1, \quad \widetilde{m}_k = m_k / \max_{1 \leqslant j \leqslant p} m_j \tag{6}$$

Here C is the color system [25], $n_c = 55$ is the number of hues, and p is the total number of pixels in an image. For an optical flow image, various colors indicate different directions of pixel movement, and the intensity of the colors indicates the magnitude of the motion. Figure 2 shows some onset frames, apex frames, and optical flow images for different micro-expressions. It can be seen that the color change in areas where the micro-expression appears is more dramatic.

3.2 Multi-scale joint feature network

In this section, we introduce the details of the proposed network structure for micro-expression

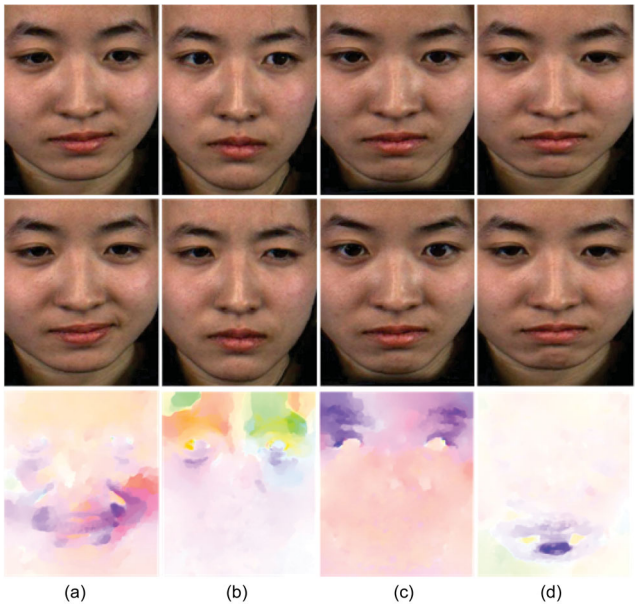

Fig. 2 Examples of the onset frame (above) and the apex frame (middle) in CASME II, and the corresponding optical flow image (below), for (a) happiness, (b) disgust, (c) surprise, (d) repression.

recognition. The structure consists of three parts: the backbone network, the joint feature module, and the fusion module. The backbone network and the joint feature module are used to extract features from optical flow images. The fusion module combines the output of the backbone network and the joint feature module to get the final prediction according to the fusion strategy. Details are given in the following subsections.

3.2.1 Backbone network

The backbone network consists of five convolutional layers, three max-pooling layers, and three fully connected layers. Detailed parameters of the backbone network configuration are shown in Table 1. The convolution layer performs multi-stage feature extraction from the optical flow image through convolution operations; different convolution layers can extract different detail features. The max-pooling layer can not only extract the features with the largest response, but also reduce the dimensionality of the feature output and feed it to the next stage of the network. At the same time, the max-pooling layer can appropriately reduce the impact of unnecessary information on model training when the number of training samples is low. The fully connected layer integrates the features from previous layers and improves the nonlinear fitting ability of the model.

3.2.2 Joint feature module

The magnitudes of micro-expressions made by each subject differ. In the process of extracting features from the optical flow image, the top convolutional layer can better capture the details of changing features, and with deepening of the network layers, detailed features may be lost. For samples with relatively subtle facial changes in the training set, we hypothesize that making full use of features from different layers is essential to recognizing micro-expressions with different amplitudes. Therefore, we propose our joint feature module (JFM) to capture micro-expression features with different magnitudes by integrating features of different layers. The structure of the JFM is shown in Fig. 3.

Table 1 Parameters of the backbone network

Type	Filter	Stride	Output
Input	—	—	224×224×3
Conv_1	11×11	4	54×54×48
MaxPooling_1	3×3	2	26×26×48
Conv_2	5×5	1	26×26×128
MaxPooling_2	3×3	2	12×12×128
Conv_3	3×3	1	12×12×128
Conv_4	3×3	1	12×12×128
Conv_5	3×3	1	12×12×128
MaxPooling_3	3×3	2	5×5×128
FC_1	—	—	256×1
FC_2	—	—	256×1
FC_3	—	—	3×1

Let $\mathcal{X}^l = \{x_i^l\}_{i=1}^c$ be the activation maps of a selected layer $l \in \{1, \ldots, 5\}$, where x_i^l is an output activation feature map of each channel i in layer l. In this work, the 2nd, 3rd, and 4th layers of the network are selected for fusion with the 5th layer. The purpose of this step is that when detailed features are lost in the feature extraction process, the fused features extracted from different network layers can still be effectively classified. Before fusion, two feature maps from different layers are linearly transformed using 1×1 convolutions to unify their dimensions, and the transformed feature maps can be formulated as follows:

$$\mathcal{X}^{l'} = \{x_j^{l'} | x_j^{l'} = \sum_{i=1}^{c} x_i^l w_{i,j} + b_j, j = 1, \ldots, c'\} \quad (7)$$

where c and c' denote the number of channels before and after convolution, respectively.

Then, the convolution results of $\mathcal{X}^l$ and $\mathcal{X}^5$ are fused to get the feature map $\mathcal{X}^m \in \mathbb{R}^{H_5 \times W_5 \times C_5}$:

$$\mathcal{X}^m = \mathcal{X}^{l'} \oplus \mathcal{X}^{5'} \quad (8)$$

where $\oplus$ denotes addition of corresponding channels.

Next, the ReLU: $\sigma(x_{c,p}^m) = \max(0, x_{c,p}^m)$ is employed to activate the feature map to obtain $\hat{\mathcal{X}}^m = \{\hat{x}_i^m\}_{i=1}^{C_5}$, to reduce interdependence between parameters and alleviate the problem of overfitting. Then, the activated feature map is input to the max-pooling layer to reduce the feature dimensionality and extract the more representative feature $\phi^m \in \mathbb{R}^{H_m \times W_m \times C_5}$,

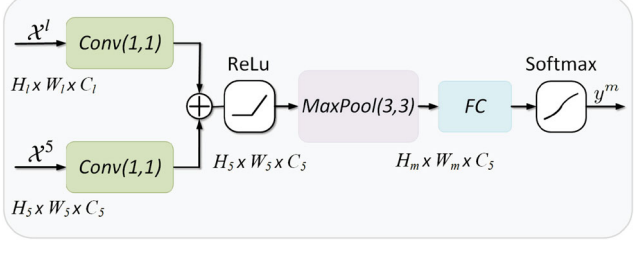

Fig. 3 Joint feature module (JFM). Combining features from different layers to obtain the prediction y^m from the module, the lower-layer features in the network can provide effective supplementing that from the higher layers.

and the subsequent fully connected layer is used to increase the non-linear fitting ability of the JFM. Finally, the softmax function normalizes the output of the fully connected layer to obtain the prediction y_j^m of the jth class, which can be expressed as follows:

$$y_j^m = \frac{\exp(W_j f(\phi^m))}{\sum_{k=1}^{N} \exp(W_k f(\phi^m))} \quad (9)$$

where $f(\phi^m)$ is the output of the fully connected layer, W is the learnable weight parameter vector, and N is the total number of micro-expression classes.

3.2.3 Fusion strategy

The prediction results of the three JFMs and the backbone network are fused to improve the recognition ability of the model. In our work, we adopt two fusion strategies to deliver the prediction results of the model.

The first strategy averages the output of the three JFMs (y^m) and the output of the backbone network (y^b) as the final prediction y :

$$y = \frac{1}{n+1} Z, \quad Z = y^{m_1} \oplus y^{m_2} \oplus y^{m_3} \oplus y^b \quad (10)$$

where n is the number of JFMs.

The second strategy uses softmax to normalize Z as the final prediction y. The probability of the jth class can be expressed as

$$y_j = \frac{\exp(Z_j)}{\sum_{k=1}^{N} \exp(Z_k)} \quad (11)$$

where j and N are the corresponding class index and the total number of micro-expression classes, respectively.

During the training phase, the cross-entropy loss function is adopted:

$$L = -\sum_{k=1}^{N} \hat{y}_k \ln(y_k) \quad (12)$$

where $\hat{y}_k$ and y_k represent the ground-truth label and prediction result for the kth sample, respectively. Through backpropagation, the backbone network and the joint feature module jointly learn to improve network performance.

4 Experiments

In this section, we first introduce the datasets adopted to evaluate the performance of the proposed approach. Secondly, we introduce the metrics and implementation details used in our experiments. Finally, we show the results of experiments, and compare and discuss the results.

4.1 Datasets

We utilized four representative datasets: SMIC, CASME II, SAMM, and 3DB, after referring to MEGC [40]. Table 2 shows the number of samples used in each dataset in our experiment. Each dataset is briefly described below.

SMIC [15] collects 164 micro-expressions from 16 participants recorded by 100 fps cameras. All types of micro-expressions are divided into three classes: positive, negative, and surprise. These micro-expression samples are only encoded using the onset frame and apex frame. Therefore, in the experiment, we calculate the average value of the optical flow between each frame and the onset frame, and the frame with the largest average value is regarded as the apex frame.

CASME II [17] contains 255 spontaneous micro-expression samples recorded by 200 fps cameras from 26 subjects, and selected from nearly 2500 induced facial movements. These samples are encoded with onset, apex, and offset frames and labeled with AUs and emotion classes. Taking MEGC as a reference, we use 145 micro-expression samples in our work. The micro-expressions are divided into three classes: positive (happiness), negative (disgust, repression), and surprise.

SAMM [18] contains 159 micro-expression samples recorded by 200 fps cameras, from 32 participants of 13 different ethnicities. The average age of participants is 33.24 years old, with equal gender ratio. These samples are encoded with onset, apex, and offset frames and labeled with AUs and emotion classes. Taking MEGC as a reference, we use 133 micro-expression samples in our experiment, and divide them into three classes: positive (happiness), negative (anger, disgust, contempt, sadness, and fear), and surprise.

3DB [40] is a combined dataset, whose samples come from SMIC, CASME II, and SAMM. The combined dataset has 68 subjects (16 from SMIC, 24 from CASME II, 28 from SAMM), so the dataset contain subjects from different backgrounds

Table 2 Samples in each dataset

Expression	SMIC	CASME II	SAMM	3DB
Positive	51	32	26	109
Negative	70	88	92	250
Surprise	43	25	15	83
Total	164	145	133	442

(ethnicity, environment, and gender). The types of micro-expression are divided into three classes: positive, negative, and surprise.

4.2 Metrics

It can be seen from Table 2 that the number of samples in different classes in each dataset is imbalanced. For example, the ratio of the three classes in the 3DB dataset is 1.3 (positive) : 3 (negative) : 1 (surprise), so accuracy cannot fully measure network performance. Therefore, we use two balanced metrics (unweighted F1-score, unweighted average recall) [42] to evaluate the performance of our method.

The unweighted F1 score (UF1) first calculates the F1 of each class, then superimposes the F1 of each class, and finally takes the average of the superimposed results according to the number of classes, which can be formulated as follows:

$$\begin{aligned} F1_c &= \frac{2TP_c}{2TP_c + FP_c + FN_c} \\ UF1 &= \frac{1}{N}\sum_{c=1}^{N} F1_c \end{aligned} \tag{13}$$

where TP_c, FP_c, and FN_c represent the number of true positives, false positives, and false negatives for the cth class, respectively. N is the number of classes.

The unweighted average recall rate (UAR) is obtained by summing the accuracy of each class, and then averaging over the number of classes:

$$UAR = \frac{1}{N}\sum_{c=1}^{N} \frac{TP_c}{M_c} \tag{14}$$

where M_c is the number of samples in the cth class.

To avoid the problem of person dependence in the classification process, we adopt leave-one-subject-out cross validation (LOSO). All facial micro-expression data of one subject are set aside for testing, and the rest are used for training to ensure that the training stage excludes the testers' information. Each subject in the dataset is then tested once.

4.3 Implementation details

To reduce the occurrence of overfitting, we applied a dropout operation on the fully connected layer and set its ratio to 0.5. In the training stage, we set the batch size to 12 and the network is trained for 300 epochs. We initialize the learning rate to 10^{-4} and use a method for stochastic optimization (Adam) to update the weights of the network. Adam is an optimization algorithm using adaptive learning rate gradient descent, which enables the network to converge faster. All of our experiments were based on an Ubuntu 16.04 system with an Nvidia GeForce GTX 1060 GPU.

4.4 Ablation study

To verify the effectiveness of JFM, we deleted the JFM module while keeping other components in the network architecture unchanged, i.e., we only used the backbone network for feature extraction and classification. The performance of the backbone network on each dataset is shown as MJFN-BbN in Table 3. We can see that when all JFMs are removed, the performance of the network declines to different degrees on each dataset, demonstrating

Table 3 Comparative results for different methods and different datasets. "—" indicates no results were reported

Method	3DB		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [40]	58.82%	57.85%	20.00%	52.80%	70.26%	74.29%	39.54%	41.02%
Bi-WOOF [7]	62.96%	62.27%	57.27%	58.29 %	78.05%	80.26%	52.11%	51.39%
MDMO [8]	—	—	62.03%	61.57%	71.72%	70.97%	—	—
Sparse MDMO [26]	—	—	67.17%	67.39%	75.98%	73.96%	—	—
AlexNet [41]	75.96%	71.98%	72.88%	71.86%	85.64%	83.77%	77.27%	66.11%
Micro-Attention [31]	50.80%	49.30%	47.30%	46.60%	53.90%	51.70%	40.30%	34.00%
ATNet [35]	63.10%	61.30%	55.30%	54.30%	79.80%	77.50%	49.60%	48.20%
OFF-ApexNet [12]	71.96%	70.96%	68.17%	66.95%	87.64%	86.81%	54.09%	53.92%
STSTNet [38]	73.53%	76.05%	68.01%	70.13%	83.82%	86.86%	65.88%	68.10%
Dual-Inception [37]	73.22%	72.78%	66.45%	67.26%	86.21%	85.60%	58.68%	56.63%
CapsuleNet [34]	65.20%	65.06%	58.20%	58.77%	70.68%	70.18%	58.82%	59.89%
Neural-Recognizer [36]	78.85%	78.24%	74.61%	75.30%	82.93%	82.09%	77.54%	71.52%
MJFN-BbN	80.71%	77.02%	73.93%	72.75%	88.92%	85.58%	71.77%	64.87%
MJFN-Avg	82.96%	79.51%	74.11%	74.13%	91.51%	88.71%	73.29%	66.15%
MJFN	**83.38%**	**80.56%**	**80.78%**	**79.95%**	**91.51%**	**88.71%**	**77.64%**	**73.12%**

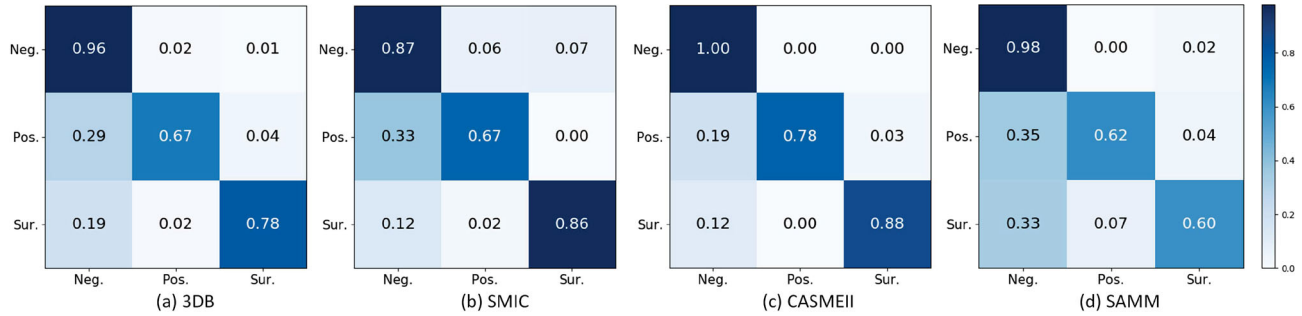

Fig. 4 Confusion matrices for the proposed method on different datasets.

the contribution of JFM to model classification performance. Using optical flow image as the input to the network in each case for fairness, a comparison to AlexNet with the same number of layers is also shown in Table 3. The backbone network (MJFN-BbN) has superior performance on the 3DB, SMIC, and CASME II datasets.

Further, we assessed the influence of the number of JFM on network performance through experiments. As can be seen from Fig. 5, with increasing number of JFM, the performance of the model improves to varying degrees on different datasets. When the number of JFM reaches 3, the overall model performance is optimal. Therefore, this work uses three JFMs. In addition, the number of JFM added is from the fourth convolution layer to the first convolution layer.

We also assessed the impact of two different fusion strategies on network performance. MJFN-Avg is the first fusion strategy, and MJFN is the second strategy. It can be seen from Table 3 that, these two strategies have the same performance on CASME II, but on the remaining datasets, the second strategy is better. Especially for the SMIC dataset, the UF1 and UAR of the second strategy are better by 6.67% and 5.82%, respectively.

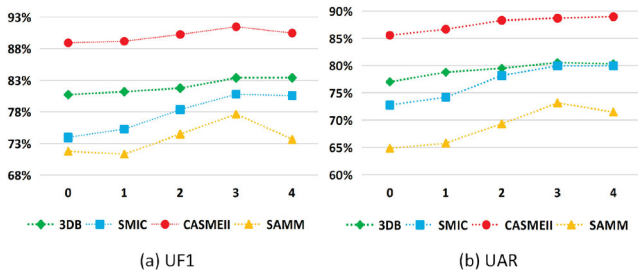

Fig. 5 Influence of the number of JFM on UF1 and UAR for different datasets.

4.5 Comparisons

We compared the proposed method with state-of-the-art methods; results are shown in Table 3. It can be seen that the proposed method (MJFN) obtains higher micro-expression recognition results. In particular, compared with the classical method LBP-TOP, UF1 and UAR of our method are better by 24.56% and 22.71%, respectively, on the 3DB dataset with extremely imbalanced samples. Compared with Bi-WOOF, which also uses apex frames, our method achieves 18.29% improvement in UAR on 3DB, while UF1 improves by 20.42%. The data shows that the performance of our proposed method on the SMIC and CASME II datasets is improved by 12% to 19% compared to the traditional methods MDMO and Sparse MDMO using optical flow. Furthermore, the proposed method also achieves 9.85% and 4.51% improvement in UF1 and UAR, respectively, compared to better results in multi-stream networks (Off-Apex, STSTNet, ATNet, Dual-Inception).

The confusion matrix in Fig. 4 shows the recognition accuracy of our proposed method on each class in each dataset. It is clear that our method has the best recognition accuracy for the negative class, which also shows that the number of samples in different classes has a certain impact on the performance of the network.

5 Conclusions

This paper proposes a novel framework for micro-expression recognition. First, we use the onset frame and the apex frame of a micro-expression sequence to generate the RGB optical flow image describing the facial changes of the micro-expression. Then, the optical flow image is fed to a multi-scale

joint feature network for feature extraction and classification. In addition, the proposed joint feature module integrates features from different layers, which helps the model to capture micro-expression features of different amplitudes. In our work, we use three JFMs to optimize the performance of the model. Furthermore, we adopt two different feature fusion strategies to fuse the prediction results of the three JFMs with the backbone network to improve the recognition ability of the model. Finally, we compare the proposed method with state-of-the-art micro-expression recognition methods; the results show that our method achieves superior results.

Although our current work can obtain superior results in micro-expression recognition, the calculation of optical flow takes a long time, which makes it difficult to guarantee real-time performance in our work. Real-time micro-expression recognition is still a major difficulty in this field. In future work, we will focus on the study of micro-expression feature extraction methods to ensure real-time micro-expression recognition.

Acknowledgements

The work was supported by the NSFC–Zhejiang Joint Fund of the Integration of Informatization and Industrialization under Grant No. U1909210, the the National Natural Science Foundation of China under Grant No. 61772312, and the Fundamental Research Funds of Shandong University (Grant No. 2018JC030).

References

[1] Haggard, E. A.; Isaacs, K. S. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In: *Methods of Research in Psychotherapy. The Century Psychology Series*. Boston: Springer, 154–165, 1966.

[2] Ekman, P.; Friesen, W. V. Nonverbal leakage and clues to deception. *Psychiatry* Vol. 32, No. 1, 88–106, 1969.

[3] Ekman, P. METT: Micro expression training tool. CD-ROM. Oakland, 2003.

[4] Huang, X. H.; Zhao, G. Y.; Hong, X. P.; Zheng, W. M.; Pietikäinen, M. Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns. *Neurocomputing* Vol. 175, 564–578, 2016.

[5] Polikovsky, S.; Kameda, Y.; Ohta, Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In: Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention, 1–6, 2009.

[6] Zhao, G. Y.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 29, No. 6, 915–928, 2007.

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

[8] Liu, Y.-J.; Zhang, J.-K.; Yan, W.-J.; Wang, S.-J.; Zhao, G.; Fu, X. A main directional mean optical ow feature for spontaneous microexpression recognition. *IEEE Transactions on Affective Computing* Vol. 7, No. 4, 299–310, 2015.

[9] Xu, F.; Zhang, J. P.; Wang, J. Z. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing* Vol. 8, No. 2, 254–267, 2017.

[10] Khor, H. Q.; See, J.; Phan, R. C. W.; Lin, W. Y. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, 667–674, 2018.

[11] Xia, Z. Q.; Hong, X. P.; Gao, X. Y.; Feng, X. Y.; Zhao, G. Y. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia* Vol. 22, No. 3, 626–640, 2019.

[12] Gan, Y. S.; Liong, S. T.; Yau, W. C.; Huang, Y. C.; Tan, L. K. OFF-ApexNet on micro-expression recognition system. *Signal Processing: Image Communication* Vol. 74, 129–139, 2019.

[13] Khor, H.-Q.; See, J.; Liong, S.-T.; Phan, R. C.; Lin, W. Dual-stream shallow networks for facial microexpression recognition. In: Proceedings of the IEEE International Conference on Image Processing, 36–40, 2019.

[14] Shreve, M.; Godavarthy, S.; Manohar, V.; Goldgof, D.; Sarkar, S. Towards macro- and micro-expression spotting in video using strain patterns. In: Proceedings of the Workshop on Applications of Computer Vision, 1–6, 2009.

[15] 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, 1–6, 2013.

[16] Yan, W. J.; Qi, W.; Liu, Y. J.; Wang, S. J.; Fu, X. L. CASME database: A dataset of spontaneous

micro-expressions collected from neutralized faces. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 1–7, 2013.

[17] 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* Vol. 9, No. 1, e86041, 2014.

[18] Davison, A. K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M. H. SAMM: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing* Vol. 9, No. 1, 116–129, 2016.

[19] Huang, X.; Zhao, G.; Hong, X.; Pietikäinen, M.; Zheng, W. Texture description with completed local quantized patterns. In: *Image Analysis. Lecture Notes in Computer Science, Vol. 7944*. Kämäräinen, J. K.; Koskela, M. Eds. Springer Berlin Heidelberg, 1–10, 2013.

[20] 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: *Computer Vision – ACCV 2014. Lecture Notes in Computer Science, Vol. 9003*. Cremers, D.; Reid, I.; Saito, H.; Yang, M. H. Eds. Springer Cham, 525–537, 2015.

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

[22] Wang, S.-J.; Yan, W.-J.; Li, X.; Zhao, G.; Fu, X. Micro-expression recognition using dynamic textures on tensor independent color space. In: Proceedings of the 22nd International Conference on Pattern Recognition, 4678–4683, 2014.

[23] Huang, X.; Wang, S.-J.; Liu, X.; Zhao, G.; Feng, X.; Pietikäinen, M. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. *IEEE Transactions on Affective Computing* Vol. 10, No. 1, 32–47, 2017.

[24] Sun, D.; Roth, S.; Black, M. J. A quantitative analysis of current practices in optical ow estimation and the principles behind them. *International Journal of Computer Vision* Vol. 106, No. 2, 115–137, 2014.

[25] Zach, C.; Pock, T.; Bischof, H. A duality based approach for realtime TV-L1 optical flow. In: *Pattern Recognition. Lecture Notes in Computer Science, Vol. 4713*. Hamprecht, F. A.; Schnörr, C.; Jähne, B. Eds. Springer Berlin Heidelberg, 214–223, 2007.

[26] Liu, Y. J.; Li, B. J.; Lai, Y. K. Sparse MDMO: Learning a discriminative feature for micro-expression recognition. *IEEE Transactions on Affective Computing* Vol. 12, No. 1, 254–261, 2021.

[27] Peng, S.; Huang, H. B.; Chen, W. J.; Zhang, L.; Fang, W. W. More trainable inception-ResNet for face recognition. *Neurocomputing* Vol. 411, 9–19, 2020.

[28] Wang, S.; Cheng, Z.; Deng, X.; Chang, L.; Duan, F.; Lu, K. Leveraging 3D blendshape for facial expression recognition using CNN. *Science China Information Sciences* Vol. 63, No. 2, 120114, 2020.

[29] 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, 382–386, 2016.

[30] 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 13th IEEE International Conference on Automatic Face & Gesture Recognition, 657–661, 2018.

[31] Wang, C. Y.; Peng, M.; Bi, T.; Chen, T. Micro-attention for micro-expression recognition. *Neurocomputing* Vol. 410, 354–362, 2020.

[32] Wang, S. J.; Li, B. J.; Liu, Y. J.; Yan, W. J.; Ou, X. Y.; Huang, X. H.; Fu, X. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* Vol. 312, 251–262, 2018.

[33] Wu, H. Y.; Rubinstein, M.; Shih, E.; Guttag, J.; Durand, F.; Freeman, W. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics* Vol. 31, No. 4, Article No. 65, 2012.

[34] Quang, N. V.; Chun, J.; Tokuyama, T. CapsuleNet for micro-expression recognition. In: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, 1–7, 2019.

[35] Peng, M.; Wang, C.; Bi, T.; Shi, Y.; Zhou, X.; Chen, T. A novel apex-time network for cross-dataset micro-expression recognition. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction, 1–6, 2019.

[36] Liu, Y. C.; Du, H. M.; Zheng, L.; Gedeon, T. A neural micro-expression recognizer. In: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, 1–4, 2019.

[37] Zhou, L.; Mao, Q. R.; Xue, L. Y. Dual-inception network for cross-database micro-expression recognition. In: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, 1–5, 2019.

[38] Liong, S. T.; Gan, Y. S.; See, J.; Khor, H. Q.; Huang, Y. C. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition. In: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, 1–5, 2019.

[39] Li, Y. T.; Huang, X. H.; Zhao, G. Y. Can micro-expression be recognized based on single apex frame? In: Proceedings of the 25th IEEE International Conference on Image Processing, 3094–3098, 2018.

[40] See, J.; Yap, M. H.; Li, J.; Hong, X.; Wang, S.-J. MEGC 2019 – The second facial micro-expressions grand challenge. In: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, 1–5, 2019.

[41] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* Vol. 60, No. 6, 84–90, 2017.

[42] Le Ngo, A. C.; Phan, R. C. W.; See, J. Spontaneous subtle expression recognition: Imbalanced databases and solutions. In: *Computer Vision – ACCV 2014. Lecture Notes in Computer Science, Vol. 9006*. Cremers, D.; Reid, I.; Saito, H.; Yang, M. H. Eds. Springer Cham, 33–48, 2015.

Xinyu Li received her B.S. degree from the School of Information and Electrical Engineering, Ludong University, in 2018. Currently, she is currently pursuing her master degree at the School of Software, Shandong University, Jinan, China. Her research interests include computer vision and image processing.

Guangshun Wei received his B.S. degree in computer science and technology from the College of Information Science and Engineering, University of Jinan, in 2017. He is currently pursuing his Ph.D. degree in the School of Software, Shandong University. His current research interests include machine learning, image processing, and geometric analysis.

Jie Wang received his B.S. degree from the School of Information and Electrical Engineering, Ludong University, in 2018. Currently, he is currently pursuing his master degree in the School of Software, Shandong University. His research interests include computer vision, image processing, and artificial intelligence.

Yuanfeng Zhou received his master and Ph.D. degrees from the School of Computer Science and Technology, Shandong University, in 2005 and 2009, respectively. He held a post-doctoral position with the Graphics Group, Department of Computer Science, the University of Hong Kong, from 2009 to 2011. His current research interests include geometric modeling, information visualization, and image processing.