Multimedia Tools and Applications (2021) 80:16125–16140
https://doi.org/10.1007/s11042-019-07896-4

MERTA: micro-expression recognition with ternary attentions

Bing Yang[1] · Jing Cheng[1] · Yunxiang Yang[1] · Bo Zhang[1] · Jianxin Li[2]

Received: 21 February 2019 / Revised: 20 May 2019 / Accepted: 13 June 2019 /
Published online: 24 June 2019

Abstract

Micro-expression is a spontaneous and uncontrollable way to convey emotions. It contains abundant psychological information, whose recognition has significant importance in various fields. In recent years, with the rapid development of computer vision, the research of facial expression tends to be more mature while the research of micro-expression remains a hot yet challenging topic. The main difficulties of recognizing micro-expression lay on the discriminative feature extraction process due to the extremely short-term and subtlety of micro-expression. To deal with this problem, this paper proposes a deep learning model to efficiently extract discriminative features. Our model is based on three VGGNets and one Long Short-Term Memory (LSTM). Three VGGNets are used to extract static and motive information where three types of attention mechanism are jointly integrated for more discriminative visual representations. Then, the spatial features of a micro-expression sequence are sequentially fed into an LSTM to extract spatio-temporal features and predict the micro-expression category. Our algorithm is carried out on the benchmark micro-expression dataset CASME II. Its efficiency is demonstrated by extensive ablation analysis and state-of-the-art algorithms.

Keywords Micro-expressions recognition · Attention mechanism · Convolutional neural network

✉ Bing Yang
yangbing7485370@126.com

Jing Cheng
chengjingwyw@sina.com

Yunxiang Yang
yyxsdu@126.com

Bo Zhang
edchang@126.com

Jianxin Li
lijx@act.buaa.edu.cn

[1] China Academy of Electronics and Information Technology, Beijing 100041, China

[2] Department of Computer Science & Engineering, Beihang University, Beijing 100041, China

1 Introduction

Facial expression is a non-verbal way to share and portray a person's feeling in our daily life. Generally, facial expression can be categorized into two classes: macro-expression and micro-expression. Macro-expression (also known as normal expression) that lasts for 3/4 to 2 seconds can be easily recognized with naked eye for humans. On the contrary, micro-expression is much shorter (1/25 to 1/3 seconds; the precise length definition varies [22, 40]) and more imperceptible. Another key characteristics of micro-expression is that it is spontaneous and uncontrollable. So even if someone tries to hide his/her true emotions and pretend another macro-expression, the micro-expression will reveal the true emotion [28]. Therefore, compared with macro-expression, micro-expression is usually regarded as a vital and accurate factor to detect human inner emotions. The research of micro-expression recognition has attracted lots of attention and gets a wide broad of applications in the field of public security [6], judicial criminal investigation [8] and so on. It has shown great power on preventing some potential danger threating to social security or warning earlier if different kinds of emergence happen. It can also benefit judging whether someone is telling the truth.

Despite the importance of recognizing micro-expression, it is extremely difficult for both machines and human beings due to its uncontrollability, short duration and small range of activity [7]. And due to the spontaneity of micro-expressions and the excitability of specific environments, the datasets of micro-expressions are very limited, which limits the design of efficient recognition algorithms to certain extent. For example, although deep learning-based algorithm has demonstrated great success on various computer vision field such as image interpolation [1], image/video enhancement [18, 19], video compression [20], and the similar topic of macro-expression recognition, its great power has not yet been fully exploited in the field of micro-expression recognition. A key reason for applying deep-learning-based method is the lack of large-scale datasets that enable efficient feature extraction for micro-expression.

Recently, visual attention mechanism has been proposed and successfully applied in structural prediction tasks such as visual captioning [39] and quality assessment [21]. It is based on the reasonable assumption that human vision tends to focus on selective parts rather than the whole visual scenes. By incorporating visual attentions, the deep models can learn richer and more efficient features for visual tasks. Therefore, visual attention can be considered as a feature extraction mechanism that combines contextual fixations.

Inspired by the visual attention mechanisms [11, 43] and the widely used convolutional neural networks (CNN), in this paper, we take full advantages of visual attention and design an attention-based CNN network for accurate micro-expression recognition called MERTA. In particular, three different types of attention have been used: 1) General attention embeds the static information of facial landmark. The facial expression is generally closely related to layout of landmark areas (the happiness will inevitably lead to the raise of the mouth corner). 2) Motive attention embeds the dynamic information. As micro-expression is characterized by tiny facial movement, it is beneficial to emphasize on the motive areas on the face. 3) Channel attention can be viewed as the process of selecting semantic attributes on the demand of the facial expression since a channel-wise feature is essentially a detector response map of the corresponding filter. For example, we want to predict disgust, our attention-wise attention will assign more weights on the channel-wise feature maps generated by filters according to the semantics like frown. We use VGGNet-16 [33] as the backbone to extract spatial features from original images and its optical flow and optical strain images. The features after the second fully-connected layers are then concatenated and fed into one layer of Long Short-Term Memory (LSTM) [9] and two fully-connected layers to predict the micro-expression. We evaluate the effectiveness of the proposed model on the well-known CASME II dataset. The proposed algorithm can significantly surpass the baseline model without attention by 4%. It also outperform state-of-the-art micro-expression algorithms.

The rest of the paper is organized as follows: Section 2 presents some related works. In Section 3, we introduce the proposed method with the details of the proposed attention mechanisms. In Section 4, we demonstrate the experimental results and ablation analysis. And Section 3.2 gives the conclusions.

2 Related work

2.1 Micro-expressions recognition

In recent years, micro-expression recognition has gradually become more popular and made remarkable progress. Pfisher et al. [28] proposed to use a temporal interpolation model to accurately recognize micro-expression. Xu et al. [38] used a facial dynamics map to identify and recognize micro-expression. Wang et al. [35] reduced the redundancy of local binary pattern from three orthogonal planes (LBP-TOP) by proposing the local binary patterns with six intersection points (LBP-SIX). In [25], the micro-expression is recognized using adaptive magnification of discriminative facial motion. The algorithm of Patel et al. [26] is the first one to explore the possibility of deep learning methods for micro-expression recognition tasks, they use transfer learning method of facial expression to select features and an evolutionary algorithm is extended to search for a set of optimal depth features. Borza et al. [2] used image difference to analyze motion changes in appointed frame and two classifiers are used to determine whether micro-expressions occur at appointed frame t. Li et al. [16] combined deep multiple task learning with normalized histogram of directional optical flow characteristics to detect micro-expressions. It divides the facial into regions of interest (ROIS) and integrate powerful optical flow methods with HOOF features to evaluate the direction of facial muscle movement. Kim et al. [13] proposed a method of recognizing micro-expressions by learning temporal and spatial feature representations with expression state constraints. Khor et al. [12] proposed an enriched long-term cyclic convolution network (ELRCN), which first encodes each micro-expression frame into an eigenvector through a CNN module. Then it predicts micro-expressions by transferring the eigenvector to a long-term and short-term memory (LSTM) module. Peng et al. [27] consider the subsize of micro-expressions in 2018 Micro-Expression Grand Challenge (MEGC). So they chose the routine of transfer learning methods to recognize micro-expressions using convolutional neutral networks. Mayya et al. [23] interpolated video sequences using time interpolation method (TIM). Then, deep convolution neural net (DCNN) was used to extract facial features in the general graphics processing unit (GPGPU) system supporting CUDA. In order to develop reliable deep neural networks, extensive training sets of labeled image samples are needed. However, due to the short face appearance and duration of depression, micro-expression recognition is still a challenging task.

2.2 Micro-expression dataset

Generally, it is difficult for ordinary people to identify micro-expressions, which have short duration, small range of change, few motion area and complex condition of psychology. At present, though many researches have been carried out on micro-expression recognition work, the datasets involved are very few. Some mainstream data sets are: 1) Spontaneous micro-expression corpus (SMIC) [15] and SMIC II proposed by Li et al. from University of Oulu in Finland. It is the first spontaneous micro-expression dataset built up in 2012 with 164 samples of 16 subjects. The micro-expressions are categorized into positive, negative, and surprise. 2) The USF-HD dataset [32] proposed by Shreve et al. from University of South Florida. It contains both micro-expression and macro-expression samples but the samples are generated by imitation rather than induction. 3) Chinese Academy of Sciences Micro-Expression (CASME) dataset [34], CASME II [41] and CAS(ME)^2 [30]. CASME contains 195 sequences with marked start frame, apex frame and end frame of micro-expression. And CASME II contains 247 spontaneous samples of 50 men and 50 women. Other micro-expression datasets include Polikovsky dataset [29] from University of Tsukuba in Japan, York DDT (New York Polygraph Testing) database [36], and Spontaneous Actions and Micro-Movements (SAMM) dataset [4] from University of Manchester. In summary, existing micro-expression datasets are limited in both sample numbers and expressions. It is mainly due to the strict environmental requirements to record micro-expression and difficulties in accurate labeling of micro-expressions. Therefore, a large-scale micro-expression dataset similar to ImageNet is almost impossible. This lays an inevitable obstacle for deep learning algorithms that are highly data-driven.

2.3 Attention mechanism

Attention mechanism [17, 42] was originally developed on the basis of human visual characteristics. Human visual attention mechanism is a unique brain signal processing mechanism: vision obtains the target area that needs to be focused on by scanning global images quickly, which we called the focus of attention normally. Then vision will invest more attention resources in this area to get more information and ignore other area with less key points. Although it has been researched decays ago, it becomes a hot topic in computer vision field recently due to its success on image/video caption and visual question answering. In [24], Mnih et al. introduced an attention mechanism to Recurrent Neural Network (RNN) for image classification. Xu et al. [37] proposed the first visual attention model for image caption. Yang et al. [43] refined the spatial attention with a stacked attention model. Semantic attention relies on semantic concepts to select effective features. In [44], the filters of convolutional layer are considered as semantic detectors. In [10], Hu et al. proposed a Squeeze-and-Excitation block (channel-wise attention) to adaptively recalibrate channel-wise feature responses. In [45], Zhang et al. proposed a context encoding module to leverage global scene context information. In SCA-CNN [3], the spatial attention and semantic attention are jointly applied for image caption.

3 The proposed algorithm

In this section, we describe the proposed micro-expression algorithm MERTA. The overall network structure is shown in Fig. 1. It is composed of three subnets of VGGNet-16 with attentions, whose output are concatenated and fed into a single layer of LSTM. We first give the descriptions on the backbone structure and then present the details on three attention mechanisms individually.

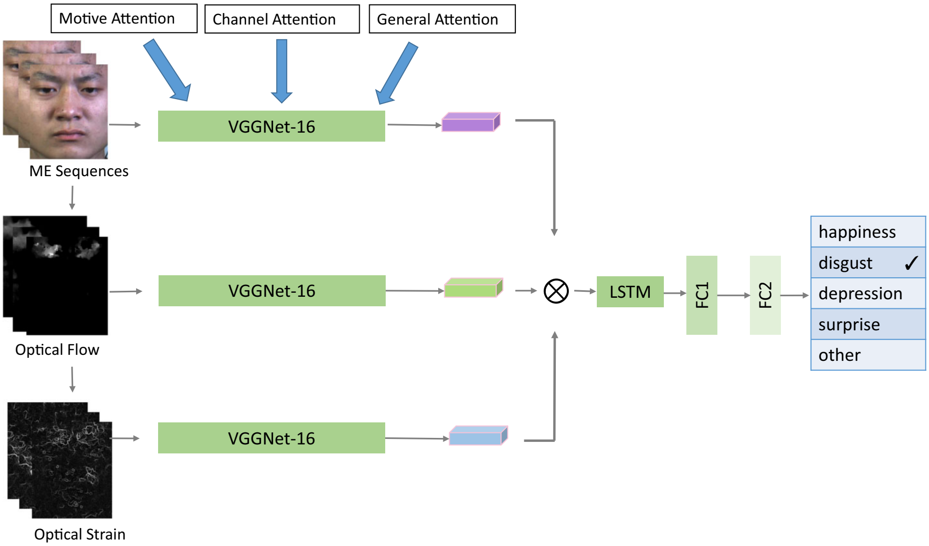


Fig. 1 Network structure of the proposed MERTA. Given a sequence of micro-expression, the input contains three parts: original frames, optical flows, and optical strains. Each part goes through a subnet of VGGNet-16. Three different attention mechanisms including general attention, motive attention, and channel attention are involved. The outputs of three subnets are then consecutively concatenated to feed into one single layer of LSTM, whose output passes two fully connected layers to predict the micro-expression categories (in this case, the micro-expression is disgust)

3.1 Backbone network

The backbone of the proposed algorithm is similar to [5] containing two parts: CNN part extracts the spatial features from each frame of the sequence and the recurrent LSTM is used to extract temporal information from consecutive spatial features. Therefore, the combination of two parts can efficiently exploit the spatio-temporal information of input sequence. Inspired by [12], we adopt enhanced inputs in our framework. In particular, the optical flow and optical strain are introduced and fed into two variants of the CNN network to extract richer hierarchical features. Optical flow captures the first-order motive information while optical strain captures higher-order derivatives, which represents the deformation incurred during non-rigid motion.

Suppose $\{I(x, y, t)\}$ is a sequence of frames, where x, y are the 2-D spatial coordinate and t represents the frame number. As a well-known motion estimation technique that is based on the brightness conservation principle, optical flow is typically defined as:

$$\frac{\partial I}{\partial x} \cdot f_x + \frac{\partial I}{\partial y} \cdot f_y + I_t = 0, \tag{1}$$

where I_t represents the temporal gradient and $\boldsymbol{f} = [f_x = \frac{\partial x}{\partial t}, f_y = \frac{\partial y}{\partial t}]$ is the optical flow. Its magnitude is denoted as $f = |\boldsymbol{f}|$. In this work, we adopt the algorithm in [31] where the optical flow is estimated using L1 norm and a regularization term. As shown in Fig. 1, the opitcal flow captures the movement of eyebrow (i.e., frown) which is closely related to the expression of disgust. According to [32], the optical strain can be calculated directly from optical flow as:

$$s = \frac{1}{2}[\nabla \boldsymbol{f} + (\nabla \boldsymbol{f})^T], \tag{2}$$

which can be expended as:

$$\boldsymbol{s} = \begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{1}{2}\frac{\partial f_x}{\partial y} + \frac{1}{2}\frac{\partial f_y}{\partial x} \\ \frac{1}{2}\frac{\partial f_y}{\partial x} + \frac{1}{2}\frac{\partial f_x}{\partial y} & \frac{\partial f_y}{\partial y} \end{bmatrix}. \tag{3}$$

Then, the magnitude of optical strain s can be computed as the L2 norm of $\boldsymbol{s}$, i.e.,

$$s = \sqrt{\left(\frac{\partial f_x}{\partial x}\right)^2 + \frac{1}{2}\left(\frac{\partial f_y}{\partial x} + \frac{\partial f_x}{\partial y}\right)^2 + \left(\frac{\partial f_y}{\partial y}\right)^2}. \tag{4}$$

As shown in Fig. 1, the optical strain captures the boundary of moving regions, highlighting the diverse deformation incurred during non-rigid facial muscle movement.

Given the original frames $\{I(x, y, t) \in \mathbb{R}^3\}$, the optical flow $\{f_x(x, y, t) \in \mathbb{R}^2\}$, $\{f_y(x, y, t) \in \mathbb{R}^2\}$, $\{f(x, y, t) \in \mathbb{R}^2\}$ and the optical strain $\{s(x, y, t) \in \mathbb{R}^2\}$, we leverage three separate classical network VGGNet-16 [33] as the backbone to fully enjoy the benefit of deep CNNs where the optical strain are first converted into a 3-channel color map by replicating it three times on channel dimension. Three types of attention mechanisms have been introduced to VGGNet-16 for further emphasized discriminative features. By extracting features from individual inputs, the separate subnets can disentangle the facial, motive, and deformable features, easing micro-expression recognition. Since high-level features generally contain semantic information, the feature maps from the second fully connected layers of three subnets are fused with concatenation and then passed to the subsequent recurrent LSTM. We follow the framework of [12] to use one single LSTM unit. But the proposed LSTM contains 256 hidden states which is less than that of [12] for compact feature representation and less memory cost. As shown in Fig. 1, a 128-d fully connected layer and a 5-d fully connected layer are appended on top of LSTM to predict the micro-expression.

Limited by the scale of micro-expression dataset, training the proposed framework end-to-end would be extremely difficult. The proposed framework is trained in two stages. In the first stage, we train three subnets of VGGNet-16 individually using micro-expression samples with labels. In order to accelerate training process and get efficient facial features with relative small-scale training samples, VGGNet-16 is initialized with the parameters of VGGFace, which is trained on the large-scale face data set Labeled Faces in the Wild (LFW). In the second stage, we fix the parameters of VGGNet-16 and train the rest layers including a LSTM module and two fully connected layers. In both of the training process, the model is optimized with a cross-entropy loss:

$$L = -\sum_k p_k \log q_k, \tag{5}$$

where k is the index for micro-expression class, p_k is one-hot vector of the ground truth micro-expression class, the q_k is the output of the softmax layer that represents the probability of different micro-expression class.

3.2 Attention mechanism

Although the three-subnet backbone has been designed to extract both static features (from original frames subnet) and dynamic features (from optical flow subnet and optical strain subnet), the features are extracted in an in discriminative way. For example, the non-facial regions are regarded equally to the facial regions. Even though this naive feature extraction method has obtained great success in image classification, face recognition or macro-expression recognition, it is far from enough for the challenging problem of micro-expression recognition due to its subtlety and shortness. As mentioned above, inspired by the smart brain signal processing mechanism, attention mechanism is shown to an effective way to emphasize on discriminative information. Different from [5] that indifferently extract facial features, the proposed algorithm introduces three types of attention mechanisms: general attention highlights the landmark areas which are fully of expression muscles; motive attention highlights the motion area where the expression appears; channel attention highlights the expression-related semantic features. The proposed model with three attention incorporated is shown in Fig. 1.

3.2.1 General attention

General attention describes the fact that all expressions including micro-expressions are more easy to be identified by movement of landmarks. For example, if someone is smiling, the most obvious symbol may be the rise of the corner of the mouth despite it takes about 42 muscles to smile. Therefore, facial landmarks are the most discriminative areas that need to concentrate. In this paper, we use dlib C++ library [14] to detect 68 landmarks of faces, $\{\boldsymbol{l}^k = [l_x^k, l_y^k], k = 1, 2, \cdots, 68\}$. As the detected landmarks are separate pixels, we further filter the landmark mask M, i.e.,

$$A_g = M * G, \tag{6}$$

where the pixels of M are all zeros expect at $\{\boldsymbol{l}^k\}$, G is the 25×25 Gaussian kernel. Then, A_g represents the general attention. The process is shown in Fig. 2. The landmarks of each frame are marked by green diamond with red numbers. It is observed that the blurry landmark highlights the critical facial region such as eyes and mouth. Emphasizing on these critical regions will facilitate micro-expression recognition.

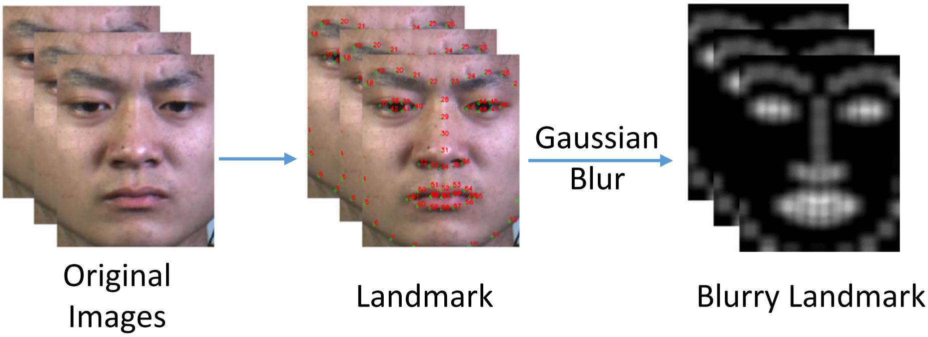


Fig. 2 General attention. We detect the landmark of each frame (marked by green diamond with red numbers) and apply Gaussian smooth process to get general attention areas

3.2.2 Motive attention

Motive attention captures the critical motive information. Since the micro-expression occurs in a very short time, it occupies only a few frames even with high-speed cameras. Therefore, trying to identify a micro-expression from few apex frames with clear spatial signal would be very difficult. Therefore, we turn to identify the motive characteristics of micro-expression. In this work, we refer to the magnitude of optical flow and optical strain for motive clues. The 2-D masks for motive attention are defined as:

$$A_m = \frac{1}{2}(f + s). \tag{7}$$

3.2.3 Channel attention

General attention and motive attention assign weights to features from the perspective of spatial dimension, which relieves the problem of distraction caused by less relevant facial regions. In fact, the same distraction problem appears on the channel dimension. As above mentioned, each feature map can be regarded as semantic responses to difference filters. Understanding and utilizing semantic information is very important for micro-expression recognition. For VGG-Face network pre-trained for face recognition, the feature maps encode rich information on appearance characteristics. Different appearance characteristics may have different levels of importance. For instance, the size of nostril will be relevant to anger than whether he/she has a hook nose. Therefore, in addition to spatial attention, we also include semantic attention in the proposed work, which is denoted as channel attention.

Given the contextual facial features extracted from conv5_3 layer of VGG-Face, our goal is to apply a set of scaling factors to automatically and selectively highlight the expression-dependent feature maps. The channel attention is shown in Fig. 3. Suppose the feature maps are represented as $\Phi = [\phi_1, \phi_2, \cdots, \phi_{512}]$, where $\phi_c \in \mathbb{R}^{W \times H}$ is the c-th slice of the feature maps Φ, 512 is the total number of channels. We first use average pooling layer to get a channel feature vector $\boldsymbol{v}$:

$$\boldsymbol{v} = [v_1, v_2, \cdots, v_{512}], \boldsymbol{v} \in \mathbb{R}^{512}, \tag{8}$$

where the average value v_c is used to represent the c-th channel features. Then two fully connected layers are exploited to learn the aggregate feature of each channel:

$$\boldsymbol{u} = \boldsymbol{W}_2 * N(\boldsymbol{W}_1 * \boldsymbol{v} + \boldsymbol{b}_1) + \boldsymbol{b}_2, \tag{9}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2$ are the convolution filters and $\boldsymbol{b}_1, \boldsymbol{b}_2$ are bias parameters. $N(\cdot)$ denotes the non-linear activation function. Note that two fully connected layers form a bottleneck

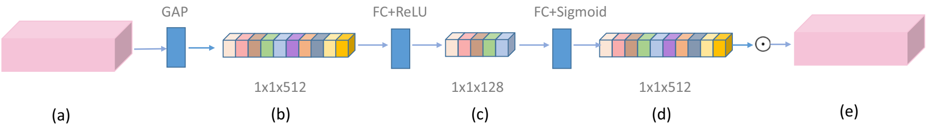


Fig. 3 Channel attention is a SE net structure to redistribute weights of different channels. First we use a pooling layer to reduce the dimension of features and put two fully-connection layers to get weights between different channels and multiple input features to achieve the redistribution operation

structure to model the correlation between channels and output the same number of weights as the input features. We first reduce the feature dimension to 1/4 of the input and then ascend back to the original dimension through a Fully Connected layer. The advantage of this method over using a Fully Connected layer directly is that it has more nonlinearity, which can better fit the complex correlation between channels and greatly reduces the amount of parameters and computation.

Then, the normalized weight vector for channel attention mechanism is then defined as:

$$A_c = \frac{1}{1 + \exp(-\boldsymbol{u})}, \tag{10}$$

which is a sigmoid function applied to $\boldsymbol{u}$. To apply the normalized weights to each channel of the input feature maps, we replicate the weight vector to the same dimension of input feature maps (i.e., $14 \times 14 \times 512$) and then perform pixel-wise multiplication.

3.2.4 Fusion of attention mechanisms

Given three attention mechanisms, their fusion procedure will obvious have great impact on the efficiency of them. Both the general attention and motive attention belong to spatial attention mechanism which guides the model to emphasize on certain spatial location of the input feature maps. General attention focuses on the landmarks while motive attention focuses on the motion area. For channel attention, it helps to emphasize on certain semantic information as each feature map can be regarded as semantic responses to different filters. Therefore, we first combine two spatial attentions (i.e., general attention and motive attention) to remove less relevant features and then apply the channel attention to emphasize on more discriminative semantic features. As shown in Fig. 4, the proposed three attention mechanisms are incorporated to the high-level feature maps from conv5_3 layer of VGGNet-16 with the dimension of $14 \times 14 \times 512$. General attention and motive attention are both with the same resolution as original frames, so they are first combined by pixel-wise summation. To match the dimension of feature maps from conv5_3 layer, the combined attention map is downsampled to 1/16 by bilinear interpolation and then replicated to the channel depth of 512. They we carry out pixel-wise multiplication between the resized attention map and feature maps. After applying spatial weights, the feature maps further go through the channel attention module to re-weight different channels of feature maps.

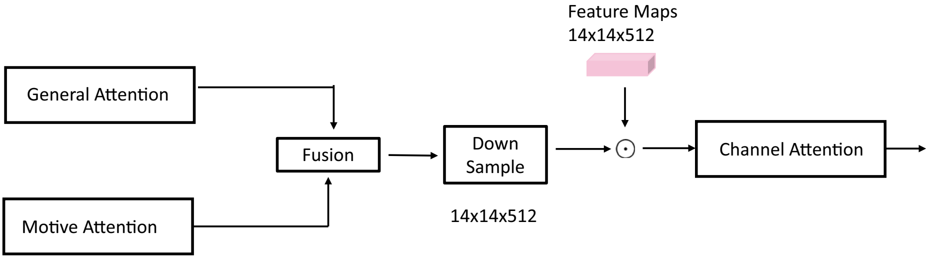


Fig. 4 The combination of three attention mechanisms. Two spatial attentions are first combined and integrated with visual features, which goes through channel attention mechanism afterwards

4 Experimental results

4.1 Experimental setting

The proposed algorithm are evaluated on the most commonly used CASME II dataset [41], which contains 255 spontaneously micro-expression video sequences recorded at high temporal resolution (200fps). The samples were divided into seven micro-expression categories, including happiness, depression, disgust, fear, sadness, surprise and others. The labels for each micro-expression was set based on not only the action unit (AU), but also the videos used to trigger emotions and responses of participants. Because the additional information may conflict in some occasions, Facial Micro-expression Challenge 1 proposed a new target class based on Facial Action Coding System (FACS), where samples are classified into seven new categories. The sample numbers of different categories vary from 1 to 99. For fair comparison, VI and VII categories are ignored. The evaluations in our experiments are conducted with leave-one-subject-out cross validation, i.e., the test subject was excluded from the training set. The recognition accuracy is then calculated by averaging 26 times evaluation (26 subjects in CASME II).

Adam optimizer is used to train our model. The initial learning rate is 10^{-5} and decay rate is 10^{-6}. We train 15 epochs to finetune three subnets of VGGNet-16 and training 20 epochs to get the final overall model. Analogy to most existing ME algorithms, the ME sequences are interpolated to fixed number of frames. In this work, Temporal Interpolation Model (TIM) [15] is used to generate 10 frames for all ME samples. As optical flow and optical strain calculate the motion between adjacent frames, there are only 9 inputs for these two subnets, and we fed only 9 ME frames into the first subnet for consistency. The ternary-attention-based visual features from 9 sets of inputs are then sequentially fed into the LSTM to get the recognition results.

4.2 Ablation analysis

To validate the efficiency of the proposed attention mechanisms, we carried out extensive ablation analysis. We compared the proposed algorithm with four different variants: Baseline model denotes the backbone network without attention mechanisms; Baseline-MA involves the motive attention; Baseline-MA-GA involved both motive attention and general attention; Baseline-MA-CA involves motive attention and channel attention; Baseline-MA-GA-CA is the proposed network with all three attention mechanisms. The accuracy results of different algorithms are shown in Table 1. It is obvious that the accuracy increases with the involvement of more attention mechanisms, validating the efficiency of each attentive component of the proposed algorithm.

Table 1 Accuracy of the proposed algorithm and its variances

	Accuracy(%)
Baseline	56.22
Baseline-MA	57.30
Baseline-MA-GA	58.38
Baseline-MA-CA	59.46
Baseline-MA-GA-CA (Proposed)	**60.54**

The best result is marked with bold

Table 2 Accuracy of the proposed and state-of-the-art algorithms

Algorithm	Accuracy(%)
LBP-TOP [28]	45.95
FDM [38]	45.93
LBP-SIP [35]	46.56
Adaptive MM + LBP-TOP [25]	51.91
ELRCN-TE [12]	52.44
Proposed	**60.54**

The best result is marked with bold

4.3 Comparison with state-of-the-art algorithms

We also compare the proposed algorithm with state-of-the-art micro-expression recognition algorithms, including: the benchmark LBP-TOP algorithm [28], the Facial Dynamics Map (FDM) algorithm [38], the modified LBP with Six Intersection Points (LBP-SIP) algorithm [35], the Adaptive Magnification of Discriminative Facial Motion (Adaptive MM + LBP-TOP) [25] and ELRCN-TE [12]. The accuracy of different algorithms are shown in Table 2. It is shown that the proposed algorithm consistently outperforms existing micro-expression algorithms with large gaps.

5 Conclusion

In this paper, we propose a novel micro-expression recognition algorithm with ternary attentions. The backbone model contains three subnet of VGGNet-16 to extract features from the original frames, the optical flow, and the optical strain, respectively. These features are then concatenated to go through one layer of LSTM for spatio-temporal features, which are used for classification with two fully connected layers. To facilitate more efficient feature extraction, we introduce three different kinds of attention mechanisms: the general attention emphasizes on the more relevant facial regions of landmarks; the motive attention guide the model to focus the facial areas with large motion; and the channel attention put more weights on the semantic features that related to micro-expressions. Experimental results validate the efficiency of each attention mechanism and the proposed model outperforms state-of-the-art algorithms with large gaps.

Acknowledgments This work was partly supported by the China Postdoctoral Science Foundation under Grant 2018M630190 and partly supported by Beijing NOVA Program (Z181100006218041).

References

1. Bao W, Lai W, Ma C, Zhang X, Gao Z, Yang M (2019) Depth-aware video frame interpolation. CoRR arXiv:1904.00830
2. Borza D, Itu R, Danescu R (2017) Real-time micro-expression detection from high speed cameras. In: IEEE International conference on intelligent computer communication and processing (ICCP), pp 357–361
3. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T (2017) SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 6298–6306

4. Davison AK, Lansley C, Costen N, Tan K, Yap MH (2018) SAMM: a spontaneous micro-facial movement dataset. IEEE Trans Affect Comput 9(1):116–129. https://doi.org/10.1109/TAFFC.2016.2573832
5. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2017) Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 39(4):677–691
6. Ekman P (2003) Darwin, deception, and facial expression. Ann N Y Acad Sci 1000(1):205–221
7. Ekman P, O'Sullivan M, Frank MG (1999) A few can catch a liar. Psychol Sci 10(3):263–266
8. Frank MG, Ekman P (1997) The ability to detect deceit generalizes across different types of high-stake lies. J Person Soc Psychol 72(6):1429
9. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
10. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 7132–7141
11. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259. https://doi.org/10.1109/34.730558
12. Khor H, See J, Phan RCW, Lin W (2018) Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: IEEE international conference on automatic face gesture recognition (FG), pp 667–674. https://doi.org/10.1109/FG.2018.00105
13. Kim DH, Baddar WJ, Yong MR (2016) Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In: ACM on multimedia, pp 382–386
14. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758
15. Li X, Pfister T, Huang X, Zhao G, Pietikäinen M (2013) A spontaneous micro-expression database: inducement, collection and baseline. In: IEEE International conference and workshops on automatic face and gesture recognition (FG), pp 1–6
16. Li X, Yu J, Zhan S (2017) Spontaneous facial micro-expression detection based on deep learning. In: IEEE International conference on signal processing (ICSP), pp 1130–1134
17. Liu J, Yang X, Zhai G, Chen CW (2016) Visual saliency model based on minimum description length. In: IEEE international symposium on circuits and systems (ISCAS), pp 990–993. https://doi.org/10.1109/ISCAS.2016.7527409
18. Liu J, Liu P, Su Y, Jing P, Yang X (2019) Spatiotemporal symmetric convolutional neural network for video bit-depth enhancement. IEEE Trans Multimed, 1–1. https://doi.org/10.1109/TMM.2019.2897909
19. Liu J, Sun W, Su Y, Jing P, Yang X (2019) BE-CALF: bit-depth enhancement by concatenating all level features of DNN. accepted by IEEE Transactions on Image Processing, pp 1–1
20. Lu G, Ouyang W, Xu D, Zhang X, Gao Z, Sun MT (2018) Deep Kalman filtering network for video compression artifact reduction. In: European conference on computer vision (ECCV), pp 568–584
21. Ma S, Liu J, Chen CW (2017) A-lamp: adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 722–731. https://doi.org/10.1109/CVPR.2017.84
22. Matsumoto D, Hwang HS (2011) Evidence for training the ability to read microexpressions of emotion. Motiv Emot 35(2):181–191
23. Mayya V, Pai RM, Pai MMM (2016) Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences. In: International conference on advances in computing, communications and informatics, pp 699–703
24. Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. In: Advances in neural information processing systems (NIPS). Curran Associates, Inc., pp 2204–2212
25. Park SY, Lee SH, Yong MR (2015) Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In: ACM international conference on multimedia, pp 911–914
26. Patel D, Hong X, Zhao G (2017) Selective deep features for micro-expression recognition. In: International conference on pattern recognition (ICPR), pp 2258–2263
27. Peng M, Wu Z, Zhang Z, Chen T (2018) From macro to micro expression recognition: deep learning on small datasets using transfer learning. In: IEEE international conference on automatic face gesture recognition (FG), pp 657–661
28. Pfister T, Li X, Zhao G, Pietikäinen M (2011) Recognising spontaneous facial micro-expressions. In: IEEE international conference on computer vision (ICCV), pp 1449–1456
29. Polikovsky S, Kameda Y, Ohta Y (2009) Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In: International conference on imaging for crime detection and prevention (ICDP), pp 1–6. https://doi.org/10.1049/ic.2009.0244
30. Qu F, Wang S, Yan W, Li H, Wu S, Fu X (2018) CAS(ME)^2: a database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Trans Affect Comput 9(4):424–436
31. Sanchez Perez J, Meinhardt-Llopis E, Facciolo G (2013) TV-L1 optical flow estimation. Image Process Line 3:137–150

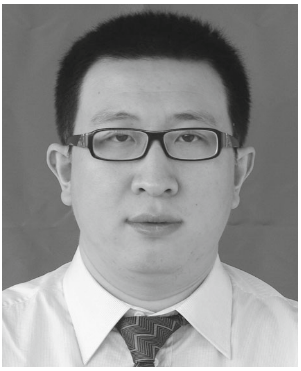
32. Shreve M, Godavarthy S, Goldgof D, Sarkar S (2011) Macro- and micro-expression spotting in long videos using spatio-temporal strain. In: IEEE international conference on automatic face gesture recognition (FG), pp 51–56
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:14091556
34. Yan W, Wu Q, Liu Y, Wang S, Fu X (2013) CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In: IEEE international conference and workshops on automatic face and gesture recognition (FG), pp 1–7
35. Wang Y, See J, Phan RCW, Oh YH (2015) LBP with six intersection points: reducing redundant information in LBP-TOP for micro-expression recognition. In: Asian conference on computer vision (ACCV). Springer International Publishing, pp 525–537
36. Warren G, Schertler E, Bull P (2009) Detecting deception from emotional and unemotional cues. J Nonverbal Behav 33(1):59–69. https://doi.org/10.1007/s10919-008-0057-7
37. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning (ICML), pp 2048–2057
38. Xu F, Zhang J, Wang JZ (2017) Microexpression identification and categorization using a facial dynamics map. IEEE Trans Affect Comput 8(2):254–267
39. Xu N, Liu AA, Liu J, Nie W, Su Y (2019) Scene graph captioner: image captioning based on structural visual representation. J Vis Commun Image Represent 58:477–485. https://doi.org/10.1016/j.jvcir.2018.12.027. http://www.sciencedirect.com/science/article/pii/S1047320318303535
40. Yan WJ, Wu Q, Liang J, Chen YH, Fu X (2013) How fast are the leaked facial expressions: the duration of micro-expressions. J Nonverbal Behav 37(4):217–230
41. Yan WJ, Li X, Wang SJ, Zhao G, Liu YJ, Chen YH, Fu X (2014) CASME II: an improved spontaneous micro-expression database and the baseline evaluation. Plos One 9(1):e86041
42. Yang B, Zhang X, Liu J, Chen L, Gao Z (2016) Principal components analysis-based visual saliency detection. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1936–1940
43. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 21–29. https://doi.org/10.1109/CVPR.2016.10
44. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision (ECCV), pp 818–833
45. Zhang H, Dana K, Shi J, Zhang Z, Wang X, Tyagi A, Agrawal A (2018) Context encoding for semantic segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 7151–7160. https://doi.org/10.1109/CVPR.2018.00747

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bing Yang received the B.E. from Wuhan University, Wuhan, China, in 2011, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2017. He is currently a postdoctoral in National Engineering Lab for Public Security Risk Perception and Control by Big Data (PSRPC) of China Academy of Electronics and Information Technology. His research interests include image and video processing, video coding and parallel realization with embedded system.

Jing Cheng received the B.Eng. and MSc. degree from Beijing University of Technology, Beijing, China, in 1987 and 1990, respectively. She is currently a researcher in China Academy of Electronics and Information Technology. His research interests include network security, information safety and data mining.

Yunxiang Yang received the Ph.D. degree in Microelectronics from Peking University, Beijing, China, in 2013. He is currently a senior engineer in National Engineering Lab for Public Security Risk Perception and Control by Big Data (PSRPC) of China Academy of Electronics and Information Technology. His research interests include computer vision, machine learning, and data mining.



Bo Zhang received the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2010. He is currently a senior engineer in National Engineering Lab for Public Security Risk Perception and Control by Big Data (PSRPC) of China Academy of Electronics and Information Technology. His research interests include machine learning, artificial intelligence and data mining.

Jianxin Li received the PhD degree from Beihang University, Beijing, China, in 2008. He is currently a professor in the School of Computer Science and Engineering, Beihang University. He was a visiting scholar in the Machine Learning Department at CMU, in 2015, and a visiting researcher at MSRA in 2011. His current research interests include data analysis and processing, distributed systems, and system virtualization.