SPATIAL-TEMPORAL ATTENTION NETWORK FOR MICRO-EXPRESSION RECOGNITION

Name of author

ABSTRACT

As a spontaneous facial muscle movement that can reveal genuine emotions, micro-expression has broad applications in national security, police interrogation, and psychological testing. It is a great challenge to recognize the micro-expression with a high accuracy for the micro-expression is a subtle facial movement. However, the existing approaches cannot solve it properly. In this paper, we proposed a simple yet effective Spatial-Temporal Attention Network (STANet). Specifically, we utilize a spatial attention module to focus on the salient regions for extracting more salient features from each frame and a temporal attention module to assign higher weights to the discriminative frames. Experiments indicate that compared with several recently published micro-expression methods, STANet achieves state-of-the-art performances on CASME I and CASME II micro-expression datasets, and competitive performance on SMIC-HS dataset.

Index Terms— Micro-expression Recognition, Deep Learning, Spatial-temporal Attention

1. INTRODUCTION

As a transient and spontaneous facial muscle movement, micro-expression reflects people's genuine emotions that cannot be hidden. Thus, micro-expression recognition has many potential applications in interrogating criminal, medical diagnosis, and commercial negotiations. Different from macro-expression, micro-expression is in low intensity, short duration, and only generates in a specific regions of a facial image [1]. Suffering from these properties, it is a great challenge to recognize human micro-expressions with a high accuracy.

In literature, many approaches have been proposed for micro-expression recognition. Among them, feature extraction methods play a crucial role on the performance of micro-expression recognition. According to the ways of feature extraction, we divide the algorithms into facial-image-based and optical-flow-based ones. Facial-image-based methods attempt to learn appearance features and extract the motion patterns of micro-expressions [2, 3]. These methods utilized spatial-temporal partition blocks to extract local features, where spatial-temporal segmentation parameters are generally used as hyper-parameters on a dataset. However, different samples may require different spatial-temporal division blocks, whereas the existing approaches use one same spatial-temporal division block for different samples. Therefore, it may cause a low accuracy in micro-expression recognition. Recently, optical-flow-based methods are proposed for micro-expression recognition [4, 5]. Since they have a good robustness in illumination changes. However, they treat the facial features as equally important whereas micro-expressions only appear in partial regions.

In recent years, the attention mechanism has become popular in computer vision domain [6], since it can focus on salient regions in an image [7] or find discriminative frames in a video [8]. By utilizing the merits of attention mechanism, we proposed a spatial-temporal attention network (STANet) based on optical-flow for micro-expression recognition in this paper, as shown in Fig. 1. Different from other spatial-temporal methods [2, 3], STANet does not need to set parameters of spatial-temporal block. Through the attention mechanism, STANet can learn more valid micro-expression regions and enhance the low intensity features of micro-expression. Furthermore, STANet can identify frames with a larger change that may include more motion information in micro-expressions. The contributions of the STANet are summarized as follows:

- We proposed a spatial-temporal attention network (STANet) for micro-expression recognition. STANet utilizes a spatial attention module to enhance the feature weights of areas of real micro-expression and a temporal attention module to learn the discriminate of each frame in the entire video clips.
- The experimental results demonstrate the effectiveness of the proposed method on several public micro-expression datasets (CASME I, CASME II, and SMIC-HS).

2. RELATED WORKS

In this section, we will briefly introduce the facial-image-based and optical-flow-based methods in micro-expression recognition. In addition, the attention mechanism will also be discussed.

Facial-image-based methods. One of the classical facial-image-based methods is local binary pattern from three orthogonal planes (LBP-TOP). It calculates LBP features from

three orthogonal planes, which are named LBP-XY, LBP-XT and LBP-YT respectively, followed by combining the features of these three planes to form the final feature vectors [2]. Based on LBP-TOP, Wang et al. [3] refined LBP with six intersection points (LBP-SIP) to reduce the redundant information of LBP-TOP. Huang et al. [9] proposed a spatio-temporal local binary pattern with integration projection (STLBP-IP), which uses image integral projection to improve the recognition performance of micro-expressions based on spatio-temporal LBP descriptor. On the basis of STLBP-IP, Huang et al. [10] further revisited the integral projection by using robust principal component analysis, followed by combining with LBP. They extracted a set of novel spatio-temporal features by incorporating shape attributes into spatio-temporal texture features. Zong et al. [11] designed a hierarchical spatial division scheme to eliminate the requirement that different micro-expression datasets must have their respective division grids. These facial-image-based methods regard spatial-temporal segmentation as hyper-parameters. However, different samples may be suitable for different spatial-temporal blocks.

Optical-flow-based methods. The optical-flow-based methods are popular for micro-expression recognition in recent years. In order to extract interpretable and intuitive information for understanding micro-expression, Xu et al. [5] divided optical flow sequences into spatio-temporal cuboids, and calculated the principal optical flow directions of each cuboid to represent the local facial dynamics (FDM). However, FDM takes a lot of time to calculate the pixel-level optical flow. At the same time, Liu et al. [4] proposed the main directional mean optical flow (MDMO) that can consider both local statistic motion information and its spatial location based on optical flow. In addition, considering the spatial changes in facial appearance are subtle, Happy et al. [12] proposed a fuzzy histogram of optical flow orientation (FHOFO) to encode the temporal pattern for classifying the micro-expressions. But the performance of FHOFO is limited. Based on MDMO, Liu et al. [13] proposed sparse MDMO to utilize a distance metric to preserve the manifold structure information of the feature space. It is well known that the valid micro-expressions only appear in partial regions of face. But these methods do not enhance the feature of regions with micro-expressions.

Attention mechanism. Attention mechanism was first proposed in the reinforce learning algorithm. Mnih et al. [6] proposed a recursive neural network (RNN)-based method to extract visual attention from videos for object recognition. Since then, attention mechanism becomes popular in many fields. As the first work for natural language processing (NLP), Bahdanau et al. [14] utilized the attention mechanism to enable translation and alignment simultaneously in machine translation. In computer vision domain, Wang et al. [7] proposed a residual attention network for image classification. Peng et al. [8] proposed a spatial-temporal attention network for video classification based on spatial and temporal coexistence relationship. Note that to the best of our knowledge, attention mechanism has not been employed in micro-expression recognition up to now.

3. SPATIAL-TEMPORAL ATTENTION NETWORK

In this section, we will introduce the details of STANet and the corresponding loss function in it. As shown in Fig. 1, STANet contains two main parts, Spatial Attention Module (SAM) which is used to extract spatial appearance features from each frame, and Temporal Attention Module (TAM) which learns the motion information from the whole sequence.

The input of STANet is an optical flow sequence $X = [x_1, x_2, ..., x_t]$, where t is the number of the optical flow frames. The pipeline of STANet is formulated as:

$$y = G(F(X)), \tag{1}$$

where F denotes the process of SAM, G denotes the process of TAM and y is the learned representation of the sequence. Specifically, the output of SAM is defined as $F(X) = [f(x_1), f(x_2), \cdots, f(x_t)]$.

3.1. Spatial Attention Module

In spatial level, we proposed a Spatial Attention Branch (SAB) to enhance the weight of regions with a higher probability of generating micro-expressions. As the key part of SAM, the SAB is a bottom-up and top-down structure, which has been introduced in pose estimation [15]. In the bottom-up phase, convolution and non-linear transformation are used to extract the appearance features of the micro-expression. And max pooling is utilized to reduce the resolution of feature map and the receptive field of feature map is gradually increased. The higher the level of feature map is, the more discriminative information could be used to identify micro-expressions. In the top-down phase, the feature map is restored to the input size using the bilinear interpolation upsampling. In order to preserve the details lost during the bottom-up phase, the original feature map is utilized by adding skip-connection [16] after each upsampling. As a result, both details and salient features can be preserved for subsequent micro-expression recognition. After the last feature map, a sigmoid layer is added to the network, and the outputs of SAB are normalized to [0,1]. With the purpose of making the network easier to converge, a residual structure is introduced [17]. As mentioned by [17], the attention module can be easily optimized to an identity map with residual structure. Thus, the performance of the network could be no worse than that without attention.

In detail, the attention mask learned by the SAB is defined as $M(X) = [m(x_1), m(x_2), ..., m(x_t)]$. For the i-th frame,

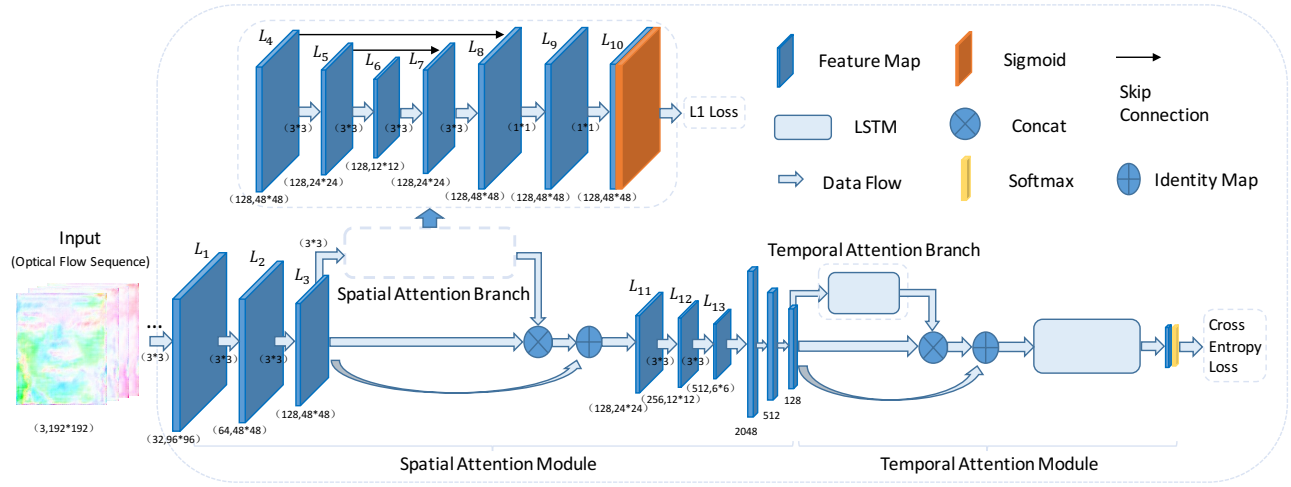

Fig. 1. STANet mainly includes two modules: 1) the Spatial Attention Module (SAM) and 2) the Temporal Attention Module (TAM). The spatial attention branch, as the key part of SAM, is a feature filter, which can enhance the salient regions of the micro-expression. Besides, the temporal attention branch is the key part of TAM and could learn a larger weight for the frames with a great variation in sequences.

the third feature map L_3 is denoted as $L_3(x_i)$, and the attention mask is $m(x_i)$. The spatial attention is expressed as :

$$S(x_i) = (1 + m(x_i)) * L_3(x_i), \tag{2}$$

where '*' is the element-wise product and $m(x_i)$ acts like a soft mask that could enhance the salient features of the micro-expression.

3.2. Temporal Attention Module

In a micro-expression sequence, the frames could be divided into the general frames (with less motion information in micro-expression) and the discriminative frames (with more motion information in micro-expression). It is crucial to recognize these discriminative frames for micro-expression recognition. Considering this factor, we propose a Temporal Attention Branch (TAB), which will pay more attention to these discriminative frames in a micro-expression sequence. As defined in Sec. 3, the input of the TAB is $F(X) = [f(x_1), f(x_2), ..., f(x_t)]$. The hidden state of $F(X)$ is denoted as a matrix H = [$\mathbf{h}_1$,$\mathbf{h}_2$,...,$\mathbf{h}_t$], where $\mathbf{h}_i$ represents the context information of the t-th time-step in the entire sequence [18]. Similar to [19], we calculate the affinity matrix C as:

$$C = \tanh(H^T H) \in \mathcal{R}^{(t \times t)}, \tag{3}$$

where C_{ij} represents the affinity coefficient between the i-th frame and the j-th frame in a micro-expression sequence. Then, the coefficient vector of entire sequence is represented as $\mathbf{w}$ = [w_1, w_2,...,w_t], where w_i represents the coefficient of the i-th frame with entire sequence and is calculated as:

$$w_i = \frac{\exp(p_i)}{\sum_{j=1}^{t} \exp(p_j)}, \quad p_i = \sum_{j=1}^{t} C_{ij}. \tag{4}$$

Moreover, the w_i is proportional to discriminant, and frames with more motion information may have a larger value. For the i-th frame, the temporal attention is expressed as $T(x_i)$ = $f(x_i)$ * w_i. In order to keep the original information, we add 1 to the normalized weights and multiply them by the original features. And the temporal attention is rewritten as:

$$T(x_i) = (1 + w_i) * f(x_i). \tag{5}$$

3.3. Loss Function

It is well known that the video classification generally utilizes the cross entropy loss function:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,j} \log(p_{i,j}), \tag{6}$$

where N denotes the number of samples, K denotes the classes of micro-expressions, $y_{i,j}$ denotes label value and $p_{i,j}$ denotes predict value. In addition, in the training phase, the output mask of the SAB does not have a clear ground-truth. However, we know that micro-expressions only generate in specific regions of facial image. As a result, the values of the mask should be sparse and only partial regions have large weights. In this paper, we regularize the mask with a l_1-weight penalty. In Sec. 3.1, the mask is defined

as $M(X) = [m(x_1), m(x_2), ..., m(x_t)]$, its corresponding l_1 norm is $||M(X)||_1$. Therefore, the total loss is denoted as:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda||M(X)||_1, \quad (7)$$

where λ is the mask penalty coefficient, controlling the importance of the mask penalty term.

4. EXPERIMENTS

Our experimental results mainly contain two parts: 1) In Sec. 4.2, STANet is compared with other state-of-the-art methods on there public datasets: CASME I [20], CASME II [21], and SMIC-HS [22]. 2) In Sec. 4.3, an ablation experiment is conducted and the effectiveness of every innovation in Sec. 3 is analyzed.

4.1. Datasets and Experimental Settings

Datasets Preparation and Experimental Metric. The CASME I dataset [20] contains 19 subjects with four kinds of micro-expressions, i.e., Disgust, Repression, Surprise and Tense. The CASME II dataset [21] contains 26 subjects with five kinds of micro-expressions, i.e., Disgust, Happy, Repression, Surprise and Others. The SMIC-HS dataset [22] contains 16 subjects with three kinds of micro-expressions, i.e., Surprise, Positive and Negative. In order to prevent overfitting, datasets are augmented during the training phase. For CASME I and CASME II, corner crop and rescaling augmentations are employed. The scaling factors are 0.9, 1.0 and 1.1 respectively. For SMIC-HS dataset, corner crop and horizontal flip augmentations are employed.

The data preprocessing are performed as follows: 1) Linear interpolation is used to interpolate all training and test samples into 20 frames. 2) Each frame is resized to 192*192 pixels. 3) Finally, for each sample , an optical flow sequence is calculated by FlowNet 2.0 [23] from adjacent frames. In our experiments, unless otherwise specified, leave-one-subject-out (LOSO) cross-validation is employed and the classification accuracy is utilized to measure the performance of the algorithm.

Settings. In all our experiments, 1) the Adam is chosen as an optimizer. 2) The learning rate is set to be 1e-5. 3) The weight decay coefficient is set as 1e-4. 4) The l_1 coefficient λ is 1e-8. 5) The model is trained for 60, 30, and 100 epochs respectively on CASME I, CASME II, and SMIC-HS datasets.

4.2. Main Results

We compare our model with state-of-the-art algorithms, including facial-image-based methods and optical-flow-based methods. The experimental results are shown in Tab. 1.

Evaluation on CASME I. It can be seen from Tab. 1 that our STANet obtains the best recognition rate. In addition,

Table 1. The comparison results with the state-of-the-art methods on CASME I, CASME II, and SMIC-HS datasets.

	CASME I	CASME II	SMIC-HS
CNN+LSTM [10]	N\A	60.98%	N\A
HSTLBP-IP [11]	N\A	63.97%	60.37%
DiSTLBP-RIP [24]	64.33%	64.78%	63.41%
HIGO [25]	N\A	67.21%	68.29%
*FHOFO [12]	N\A	N\A	51.83%
*FDM [5]	56.14%	45.93%	54.88%
*MDMO [4]	56.29%	51.69%	58.97%
*Sparse MDMO [13]	74.83%	66.95%	**70.51**%
*Ours	**76.61**%	**69.11**%	68.90%

* represents optical-flow-based methods.

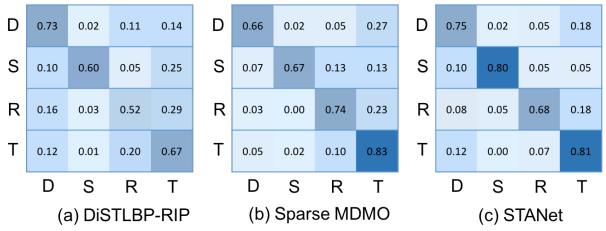

Fig. 2. The confusion matrices of DiSTLBP-RIP, Sparse MDMO, and STANet on the CASME I dataset. Abbreviation symbols, shown in the matrix, are denoted as: D (Disgust), S (Surprise), R (Repression), T (Tense).

we compare the confusion matrices of DiSTLBP-RIP, Sparse MDMO and STANet, as shown in Fig. 2. Compared with DiSTLBP-RIP and Sparse MDMO, the recognition performance on four kinds of micro-expressions is relatively close in STANet, indicating that our model is more stable in the unbalanced dataset (CASME I). Especially, the recognition rate of "Surprise" in STANet far exceeds the other two methods.

Table 2. Ablation study on CASME I dataset. l_1 stands for l_1 loss.

	CASME I
baseline	71.35%
baseline+SAB	73.10%
baseline+SAB+l_1	74.85%
baseline+SAB+TAB+l_1	76.61%

Evaluation on CASME II. STANet achieves the best recognition accuracy on the CASME II. The confusion matrices of DiSTLBP-RIP, Sparse MDMO and STANet are shown in Fig. 3. The results indicate that STANet achieves the best recognition rate in "Surprise" and "Others" micro-expressions. Surprisingly, STANet even achieves a 96% recognition rate on "Surprise". One possible reason is that the

Fig. 3. The confusion matrices of DiSTLBP-RIP, Sparse MDMO and STANet on the CASME II dataset. Abbreviation symbols, shown in the matrix, are denoted as: H (Happy), D (Disgust), S (Surprise), R (Repression), O (Others).

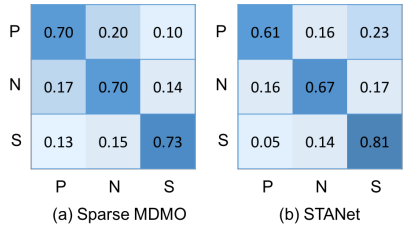

Fig. 4. The confusion matrices of Sparse MDMO and STANet on the SMIC-HS dataset. Abbreviation symbols, shown in the matrix, are denoted as: P (Positive), N (Negative), S (Surprise).

degree of change on "Surprise" is relatively large and STANet can pay more attention to those areas with great variation.

Evaluation on SMIC-HS. In addition to Sparse MDMO, STANet achieved the best recognition on the SMIC-HS. The confusion matrices of Sparse MDMO and STANet are shown in Fig. 4. The experimental results show that STANet can achieve a better recognition effect on "Surprise", but the recognition performance of "Positive" is can be improved.

4.3. Ablation study

In this subsection, ablation experiments are performed on CASME I to demonstrate the effectiveness of SAB (Sec. 3.1), TAB (Sec. 3.2) and l_1 loss (Sec. 3.3). The results are shown in Tab. 2. In addition, we visualize the spatial attention in Fig. 5 and the temporal attention in Fig. 6.

As shown in Tab. 2, the model with SAB exceeds the baseline by 1.75%. The l_1 normalization further improves the results by 1.75%. Fig. 5 visualizes the attention maps learned by SAB with l_1 norm. The Fig. 5 (a) displays a sample with label "Surprise". It can be seen that the feature map with attention focuses more on cheeks and forehead which are salient regions of "Surprise". Similarly, in Fig. 5 (b), feature map with attention focuses on mouth and nose which are salient regions of "Happy".

The last line of Tab. 2 shows that TAB brings 1.76% improvement. The visualization of temporal attention is shown

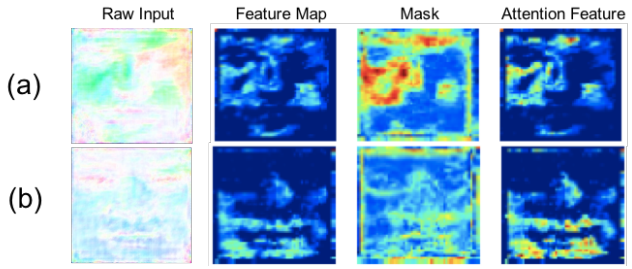

Fig. 5. The visualization of spatial attention. The sample (a) shows the "Surprise" while the sample (b) shows the "Happy". The information of the feature map will gradually increase when the color is changed from blue to green, while the information will decrease if the color is changed from green to red. Note that blue contains the fewest information than other colors.

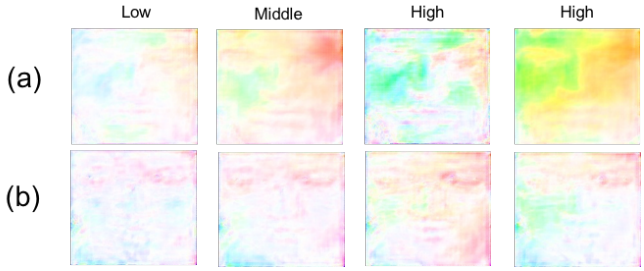

Fig. 6. The visualization of temporal attention. The expression intensity of the optical flow picture are marked on the top. The sample (a) and (b) are selected from two different micro-expression sequences. For the optical flow pictures, different colors indicate different directions of the motion, whose magnitude is usually proportional to the brightness of the color.

in Fig. 6. Empirically, the motion information contained in the optical flow frame is proportional to its discriminating ability. The weights of the four frames in sample (a) are 0.0003, 0.0372, 0.0778 and 0.0852 respectively, and the weights of the four frames in sample (b) are 0.0068, 0.0481, 0.0756 and 0.0809. This means the temporal attention does allocate higher weight to more discriminative frames.

5. CONCLUSIONS

In this paper, we proposed an attention-based network, named STANet, for the task of micro-expression recognition. Specifically, a spatial attention branch is introduced in our model to emphasize the salient regions of the micro-expression through giving more weights to these regions. Besides, a temporal attention branch is incorporated to learn the more discriminative frames by putting more weights on these discriminative frames. Experiments on three public micro-expression datasets show that compared with other state-of-the-art methods, our model achieved the highest prediction performance

on CASME I, CASME II, and competitive prediction performance on SMIC-HS. Note that by focusing on the structure of the network, the proposed STANet only utilizes a simple loss function. In the future, we will utilize more efficient loss function for our spatial-temporal micro-expression recognition model.

6. REFERENCES

[1] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, "Micro-expression recognition using color spaces," *IEEE TIP*, vol. 24, no. 12, pp. 6034–6047, 2015.

[2] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *ICCV*, 2011.

[3] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *ACCV*, 2014.

[4] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE TAFFC*, vol. 7, no. 4, pp. 299–310, 2016.

[5] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE TAFFC*, vol. 8, no. 2, pp. 254–267, 2017.

[6] V. Mnih, N. Heess, A. Graves, and K. kavukcuoglu, "Recurrent models of visual attention," in *NIPS*, 2014.

[7] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017.

[8] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE TCSVT*, 2018.

[9] X. Huang, S.-J. Wang, G. Zhao, and M. Pietikäinen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *ICCVW*, 2015.

[10] H. Xiaohua, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE TAFFC*, 2018.

[11] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE TMM*, vol. 20, no. 11, pp. 3160–3172, 2018.

[12] S. L. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE TAFFC*, 2018.

[13] Y. Liu, B. Li, and Y. Lai, "Sparse MDMO: Learning a discriminative feature for spontaneous micro-expression recognition," *IEEE TAFFC*, 2018.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[18] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *CVPR*, 2017.

[19] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS*, 2016.

[20] W.-J. Yan, Q.Wu, Y.-J. Liu, S.-J.Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *FGR*, 2013.

[21] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLOS ONE*, vol. 9, no. 1, pp. e86041, 2014.

[22] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *FGR*, 2013.

[23] I. Eddy, M. Nikolaus, S. Tonmoy, K. Margret, D. Alexey, and B. Thomas, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017.

[24] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE TAFFC*, 2018.

[25] X. Li, X. HONG, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE TAFFC*, 2018.