CROSS-DATABASE MICRO-EXPRESSION RECOGNITION: A STYLE AGGREGATED AND ATTENTION TRANSFER APPROACH

Ling Zhou, Qirong Mao, Luoyang Xue*
School of Computer Science and Communication Engineering, Jiangsu University, China
Emails: {2111808003@stmail.ujs.edu.cn, mao_qr@ujs.edu.cn, 2211708031@stmail.ujs.edu.cn}

ABSTRACT

Deep learning systems, such as the Residual Network (ResNet), can infer a hierarchical representation of input data that facilitates categorization. In this paper, we propose a Style Aggregated and Attention Transfer framework (SA-AT) based on ResNet for cross-database Micro-Expression Recognition (MER). The training of SA-AT has two stages. In the first stage, facial expression samples are used as the auxiliary database to train a ResNet teacher model. To benefit from the significant classification ability of teacher model which is trained on large scale data, in the second step, the attention of the teacher model is transferred to train the student model on style aggregated micro-expression databases with limited samples. Our experiments demonstrate that compared with the baseline method in Micro-Expression Grand Challenge 2019, our proposed technique achieves more promising performance.

Index Terms— cross-database micro-expression recognition, micro-expression recognition, transfer learning

1. INTRODUCTION

As an essential way of human emotional behavior understanding and lie detecting, in the past ten years, Micro-Expression Recognition (MER) has attracted a great deal of attention in human-centered computing. In MER, cross-database MER is a challenging and interesting research topic, which means that the training and testing samples are from different databases [1]. There are two cross domain protocols proposed by the Micro-Expression Grand Challenge (MEGC) 2018 [2]- Composite Database Evaluation (CDE) and Holdout-Database Evaluation (HDE). HDE and CDE are Tasks A and B respectively in MEGC 2018. HDE means that training and testing sets are to be sampled from different databases and evaluated; while CDE means that all samples from all used micro-expression databases are combined into a single composite database. Leave-One-Subject-Out (LOSO) cross-validation is used to determine the training-testing splits.

Some strategies of cross-database MER have been proposed, besides being categorized to CDE and HDE tasks, which can also be mainly classified into two categories: emotion-based methods [1, 3], and AUs-based methods [4, 5]. Compared with AUs-based methods, emotion-based methods may obtain more straightforward recognition results. In our work, we aim to tackle the task of emotion-based micro-expression recognition on the cross-database according to the challenge of MEGC 2019[1].

Recently, deep learning strategies, such as style aggregated method [6] and attention transfer mechanism [7, 8], have been successfully applied to the field of domain adaption and transfer learning. These strategies could reduce the domain gap between different styles of images, and infer a hierarchical feature representation for facilitating categorization, e.g., Facial Expression Recognition (FER). In this paper, to reduce the domain gap between different micro-expression databases, and transfer strong classification capability of deep learning network from the large scale data to limited simples data, we propose a Style Aggregated and Attention Transfer (SA-AT) approach to deal with the CDE task on emotion category. Specially, the facial expression database is applied as an axillary database to train the teacher model, then by transferring the attention of the teacher model, a student model learns on the style-aggregated micro-expression samples. Compared with the popular baseline method of cross-database MER method used in MEGC 2019, experimental results show that our proposed strategy acquires more promising performance. The major contributions of this paper are summarized as follows:

(1) To the best of our knowledge, this is the first study that introducing style aggregated method to cross-dataset MER field. Style aggregated method can properly bridge the domain gap between different datasets by regenerating new micro-expression images with aggregated style.

(2) Our approach handles the limited-samples cross-database MER problem, by introducing simple yet effective attention transferred mechanism and a large-scale-data FER deep learning method.

The rest of the paper is organized as follows. Section 2 presents our SA-AT algorithm in details. Section 3 describes cross-database MER benchmark datasets and our experimental results. Conclusion are discussed in Section 4.

*Corresponding Author

[1] https://facial-micro-expressiongc.github.io/MEGC2019/

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00025

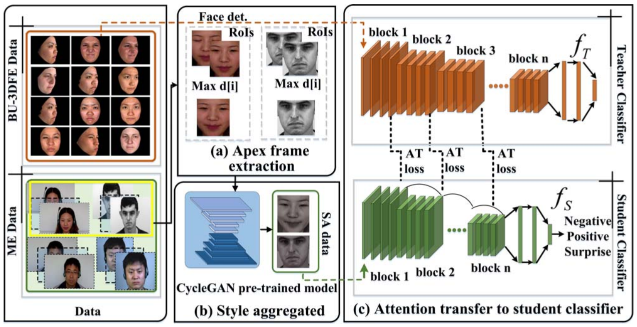


Fig. 1: The overall architecture of the proposed approach of SA-AT, which consists of three important modules: Apex frame extraction, style-aggregated and attention transferred.

2. SA-AT

In the pre-processing step of the proposed approach, we extract apex frames from micro-expression videos and transferring all the micro-expression images to an aggregated style, and then we learn a deep learning teacher model from facial expression images. After that, micro-expression samples are used to learn the student model with a teacher model's transferred attention. The overview of our proposed approach is shown in Fig. 1. In the following subsections, we will elaborate on the three important components of the framework in detail.

2.1. Apex Frame Extraction

The apex frames of micro-expression videos are selected by automatic apex frame spotting strategy [9]. The procedure of extracting the apex frames is shown in Algorithm 1. By Landmark detection and regions-of-interest (RoIs) extraction of every frame in micro-expression videos, the differences between the first frame and the rest i-th frame is defined as Eq. (1):

$$d(i) = 1 - \frac{\sum_{j=1}^{nBins} h_{1j} \times h_{ij}}{\sqrt{\sum_{j=1}^{nBins} h_{1j}^{i} \times \sum_{j=1}^{nBins} h_{ij}^{2}}}, \tag{1}$$

where h_{1j} is the gray-scale histogram of the first frame, and h_{ij} is the i-th frame in a micro-expression video.

2.2. Samples Style Aggregated

Style aggregated model aims to transform faces with different styles into an aggregated style. In our methodology, we aggregated the micro-expression samples from different micro-expression datasets using the pre-trained style aggregated model of [6]. The principle of the algorithm is to combine a large number of different styles of images together to generate a style-aggregated output via CycleGAN [10].

Algorithm 1 Apex Frame Extraction

```
Inputs: $X(n) = (x[1], x(2), .., x[n])$: $n$ frames in a micro-expression ideo clip;
Outputs: $x[af\_num]$: the apex frame of the micro-expression sample;
$maxd = 0$;
for $i = 1; i < n + 1; i + +$ do
    detect 68 landmark points $P(i)$ by lib face detection;
    extract three RoIs according to the 68 landmark points;
    get LBP histograms in RoIs;
    if $i > 1$ then
        calculate $d[i]$ as Eq. (1);
        if $d[i] > maxd$ then
            $af\_num = i$;
            $maxd = d[i]$;
        end if
    end if
end for
return $x[af\_num]$
```

2.3. Attention Transfer

The basic idea of attention transfer is as follows. Considering we have a trained Convolution Neural Network (CNN) model and a test image x, a target expression c, and the corresponding J feature maps A^j of a CNN layer. The image x is first forwardly propagated through the trained CNN model, and then a spatial attention map is constructed by computing statistics of the feature maps across all the J channel dimensions:

$$F(A) = \sum_{j=1}^{J} |A^j|. \tag{2}$$

The key focus of the attention transfer is to learn knowledge transfer from a deeper network to a shallower network within the same classification or recognition goal.

Teacher classifier learning: We start with learning a teacher classifier f_T that can perform the classification task on the facial expression images X_T with labels Y_T belong to c categories. Here we adopt residual network (ResNet) following the architecture of [11], taking X_T as inputs. We use typical softmax cross-entropy loss for the teacher classifier T. which corresponds to:

$$\begin{aligned} \mathcal{L}_T\left(f_T, X_T, Y_T\right) &= E_{(x_t, y_t) \sim (X_T, Y_T)} \\ &- \sum_{c=1}^{C} 1[c = y_t] \log\left(\sigma\left(f_T^{(c)}\left(x_t\right)\right)\right), \end{aligned} \tag{3}$$

where σ denotes the softmax function, while the learned model f_T will perform well on the auxiliary facial expression data, then the student classifier can benefit from the strong classification ability of it.

Student classifier learning: To take the advantages of the method of FER in teacher model learning, we train the student classifier f_S on micro-expression data by distilling knowledge from f_T and here we do so by attention transfer which has been proved in [7]. In attention transfer module, the goal is to train a student network that will not only make correct predictions but will also have attentions maps that are similar to those of the teacher. This corresponds to the loss function:

$$\begin{aligned} \min \mathcal{L}_S(f_S, X_S, Y_S) &= E_{(x_s,y_s)\sim(X_S,Y_S)} \\ &-\sum_{c=1}^{C} 1[c = y_i^t] \log\left(\sigma\left(f_S^{(c)}(x_s)\right)\right) \\ &+\frac{\beta}{2}\sum_{j\in I}\left\| \frac{Q_S^j}{\left\|Q_S^j\right\|_2} - \frac{Q_T^j}{\left\|Q_T^j\right\|_2} \right\|_2, \end{aligned} \tag{4}$$

where the first item is a typical softmax cross-entropy loss, which is the same as Eq. (3). The second item is the attention transfer loss where $Q_S^j = vec(F(A_S^j))$ and $Q_T^j = vec(F(A_T^j))$ are respectively the j-th attention maps pair of the classifier f_S and f_T in vectorized form, and $F(.)$ is calculated according to Eq. (2). β is the weight of the attention transfer loss. Based on Eq. (4), the cross-database MER with limited training samples is implemented.

3. EXPERIMENTS & RESULTS

3.1. Datasets

Three widely-adopted public micro-expression databases are employed for our experimental evaluations: SMIC [12], CASME II [13] and SAMM [14]. These databases are also benchmark databases in MEGC 2019.

SMIC [12]: There are 164 micro-expression clips from 16 different subjects at 100 *fps* in SMIC, with 3 ethnicities. The resolution of samples is 640×480 *pixels*. There are three micro-expression types in SMIC, including *Negative*, *Positive* and *Surprise*.

CASME II [13]: The CASME II dataset contains 256 micro-expressions samples from 26 subjects at 200 *fps*. CASME II dataset includes only Chinese participants. The resolution of the samples are 640×480 *pixels*. The samples in CASME II are categorized into five micro-expression classes, including *Happiness*, *Surprise*, *Disgust*, *Repression* and *Others*.

SAMM [14]: The SAMM dataset contains 159 micro-expression instances from 32 participants at 200 *fps*. These participants are from 13 races and the resolution of the samples are 2040×1088 *pixels*. The samples in SAMM demonstrates seven micro-expression classes including *Happiness*, *Surprise*, *Disgust*, *Repression*, *Angry*, *Fear* and *Contempt*.

3.2. Experiment setup

We evaluate our model on the same experimental settings as the micro-expression recognition challenge of MEGC 2019.

Table 1: The sample information of expression datasets.

Dataset	Expression Category			
	Negative	Positive	Surprise	Total
BU-3DFE [15]	11012	2730	2736	16478
SMIC [12]	70	51	43	164
CASME II [13]	88	32	25	145
SAMM [14]	92	26	15	133

(1) Three main emotion categories: To facilitate classification based on common grouping of expression, the original emotion classes are regrouped into three main emotions. Specifically, the labels of *Surprise* samples are unchanged, the samples of *Happiness* are relabeled as *Positive*. The samples of *Disgust*, *Repression*, *Anger*, *Contempt*, *Fear* and *Sadness* are grouped into *Negative* class. The detail information of the composite and three individual databases of SMIC, CASME II, and SAMM are shown in Table 1. After relabeling the samples, there are altogether 68 subjects (16 from SMIC, 24 from CASME II, 28 from SAMM) in the composite database. As shown in Table 1, there are total of 442 samples in the composite database, and the class imbalance exists in almost all the databases.

(2) LOSO cross-validation: Leave-One-Subject-Out (LOSO) cross-validation means that in all the folds of experiments, each subject group is held out as the testing set while all remaining samples are used for training. Therefore, there are 68 LOSO folds (total of 68 subjects in the composite database) needed for one round experiment.

(3) UAR and UF1 metrics: To properly handle such class imbalances [16], two well-accepted balanced metrics of cross-database MER are recommended: Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) to measure the performance of various methods. The results for the composite database, and the individual SMIC, CASME II and SAMM parts should be reported. These metrics are calculated as follows:

$$UF1 = \frac{1}{C}F1_c, \tag{5}$$

and

$$UAR = \frac{1}{C}\sum Acc_c. \tag{6}$$

Such that,

$$F1_c := \frac{2 \cdot \sum_{k=1}^{K} TP_c^{(k)}}{2 \cdot \sum_{k=1}^{K} TP_c^{(k)} + \sum_{k=1}^{K} FP_c^{(k)} + \sum_{k=1}^{K} FN_c^{(k)}}, \tag{7}$$

and

$$Acc_c = \frac{\sum_{k=1}^{K} TP_c^{(k)}}{n_c}, \tag{8}$$

where n_c is the total number of samples in the ground truth of the class c, For the k-th (of K folds) on LOSO by each class c

Table 2: Results of facial expression recognition methods.

Dataset	Method	WAR	UAR	UF1
BU-3DFE [15]	AlexNet	0.7518	0.7215	0.7522
	VGGNet-16	0.7642	0.7432	0.7748
	ResNet-38	0.8211	0.7615	0.7865
	ResNet-50	0.8441	0.7779	0.7957
	ResNet-110	**0.8968**	**0.8634**	**0.8952**

(of C classes), $TP_c^{(k)}$ is the true positives, $FP_c^{(k)}$ is the false positives and $FN_c^{(k)}$ is the false negatives. To get the UF1 and UAR, it first needs to compute each c-th class respective F1-scores $F1_c$ and accuracy Acc_c.

All the experiments are conducted with Ubuntu 16.04, Python 3.6.2 with Keras 2.2.4 and Tensorflow 1.11.0 on an NVIDIA GTX Titan GPU with 16 GB on-board memory. The batch size is set to 64 and 16 for teacher model and student model learning, respectively. Learning rate of teacher and student models are 0.0001. All weights are initialized from a zero-centered normal distribution with a standard deviation of 0.02.

3.3. Preprocessing & Settings

The network is constructed as shown in Fig. 1. We first select the apex frames to pre-process the micro-expression samples. For all the image samples, we use the lib face detection algorithm [17] to crop out the faces and resize them as 256 × 256.

Then we compare our model with the baseline method (LBP-TOP) [18, 19]. Specifically, in the baseline method, Temporal Interpolation Model (TIM) [12] of length 10 is applied. The neighboring radius R on the three orthogonal planes are set as 1, 1 and 4, respectively. And the number of the neighboring points P for all planes are set as 4. Besides, 5 × 5 grids are adopted to partition the micro-expression video clips into a few facial blocks without any overlapping blocks.

3.4. Teacher Model Selection

In order to choose a quantification teacher model for our task, we investigate different basic facial expression classification models (AlexNet [20] and ResNet [11]) on BU-3DFE [15]. As shown in Table 2, compared with the others, the performance of ResNet-110 is better. Therefore, in our experiment, the basic classification model is fixed as ResNet-110.

3.5. Performance Evaluation

To help analyze our SA-AT model and show the benefit of each module, we designed several experimental methods as follows:

(1) **ResNet:** This baseline uses the BU-3DFE dataset as the auxiliary database to pre-train the teacher model of ResNet110 and then tests it on original micro-expression samples.

Table 3: Comparisons among the evaluated modules.

Method	AT	SA	UF1	UAR
ResNet			0.5415	0.5348
ResNet&SA		✔	0.5611	0.5503
ResNet&AT	✔		0.5782	0.5794
SA-AT	✔	✔	**0.5936**	**0.5958**

(2) **ResNet with style aggregated (ResNet&SA):** This baseline differs from the ResNet that using style aggregated micro-expression samples for testing instead of original samples. By comparing it with the ResNet, we can evaluate the effect of style aggregated module.

(3) **ResNet with attention transfer (ResNet&AT):** This baseline adds attention transfer to the Resnet. Specially, we first pre-train the teacher model on the BU-3DFE, then train the student model on the original micro-expression data by the attention transfer mechanism. By comparing it with the Resnet, we can evaluate the effect of attention transfer module.

(4) **ResNet with style aggregated and attention transfer (SA-AT):** This module adds style aggregated to the ResNet&AT. Different from the ResNet&AT, all the samples used in the SA-AT are generated by the style aggregated module. By comparing it with the ResNet&AT, we can evaluate the effect of style aggregated strategy co-occurrences with attention transfer.

In Table 3 , we show the differences between the above modules. The detail comparison results are illustrated as follows.

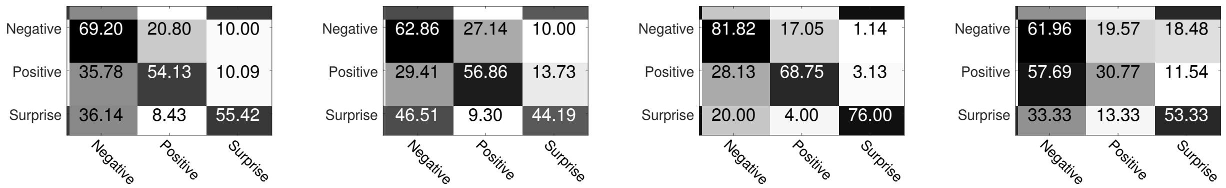
Comparison results on the evaluated modules: These results in Table 3 clearly show that performance on the models with attention transfer and style aggregated are better than those without them, which indicates that the two strategies: attention transfer and style aggregated are made the contribution to the improvement of recognition performance.

Comparison with the baseline method: We compare our model with the baseline method of LBP-TOP in the MEGC 2019. As shown in Table 4, clearly, our method SA-AT can achieve higher UF1 and UAR than the baseline. This may attribute to the two modules which mentioned in Section 2: (1) Aggregated of samples styles which can reduce the domain shift between different datasets; (2) Attention transfer which can benefit from the strong category capability of deep learning network based on large scale data.

Furthermore, it can be found that the UF1 and UAR in CASME II is highest among the three databases when we utilize the SA-AT model. It indicates that compared with the database of CASME II, the recognition in the databases of SAMM and SMIC are more challenging. For SAMM dataset, it is very likely due to the higher class imbalance problem and

Table 4: Performance metrics comparison among the baseline of MEGC 2019 and our proposed method on the composite database and the individual databases.

Methods	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP[18]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
SA-AT(Ours)	**0.5936**	**0.5958**	**0.5512**	**0.5463**	**0.7607**	**0.7552**	**0.4476**	**0.4868**



(a) Full of SA-AT

(b) SMIC of SA-AT

(c) CASME II of SA-AT

(d) SAMM of SA-AT

Fig. 2: The confusion matrices of SA-AT composite and individual databases, respectively.

the higher variance in the race of the participants. For the S-MIC dataset, the lower resolution of the face area and frame rate mostly contribute to the lower performance.

Confusion matrices of the SA-AT model: The confusion matrices of SA-AT on the combined database and the individual databases are depicted in Fig. 2. To check the above analysis and further to indicate that our experimental results are achieved from the composite database. It can be observed that excluding the correct predicted samples, most of the other *Positive* and *Surprise* samples are predicted to *Negative* category. This mainly due to the class imbalance occurring in the task and *Negative* micro-expression class is the dominant category in the databases.

4. CONCLUSION

In this paper, we propose a deep learning framework for cross-database MER. With style aggregated and attention transfer strategy, the proposed model can categorize micro-expression emotions effectively on the three public micro-expression datasets. Experiments show the superior performance of our model when comparing with the baseline method of MEGC 2019. As future work, more detailed categories such as five micro-expressions based experiments will be investigated.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under grant Nos. 61672267 and U1836220.

References

[1] Yuan Zong, Xiaohua Huang, Wenming Zheng, Zhen Cui, and Guoying Zhao, "Learning a target sample re-generator for cross-database micro-expression recognition," in *Proceedings of the 25th ACM International Conference on Multimedia (ACM MM)*, 2017, pp. 872–880.

[2] Moi Hoon Yap, John See, Xiaopeng Hong, and Su-Jing Wang, "Facial micro-expressions grand challenge 2018 summary," in *Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018, pp. 675–678.

[3] Yuan Zong, Wenming Zheng, Xiaohua Huang, Jingang Shi, Zhen Cui, and Guoying Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Trans. Image Processing*, vol. 27, no. 5, pp. 2484–2498, 2018.

[4] Huai-Qian Khor, John See, Raphael Chung-Wei Phan, and Weiyao Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018, pp. 667–674.

[5] Adrian K. Davison, Walied Merghani, and Moi Hoon Yap, "Objective classes for micro-facial expression recognition," *J. Imaging*, vol. 4, no. 10, pp. 119, 2018.

[6] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, "Style aggregated network for facial landmark detection," in *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp.

379–388.

[7] Sergey Zagoruyko and Nikos Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” *CoRR*, vol. abs/1612.03928, 2016.

[8] Songsong Wu, Xiao-Yuan Jing, Dong Yue, Jian Zhang, K. Jian Yang, and Jingyu Yang, “Unsupervised visual domain adaptation via dictionary evolution,” in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[9] Sze-Teng Liong, John See, KokSheik Wong, Anh Cat Le Ngo, Yee-Hui Oh, and Raphael C.-W. Phan, “Automatic apex frame spotting in micro-expression database,” in *Proceedings of 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 665–669.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen, “A spontaneous micro-expression database: Inducement, collection and baseline,” in *Proceedings of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[13] Wenjing Yan, Xiaobai Li, Sujing Wang, Guoying Zhao, Yongjin Liu, YH Chen, and Xiaolan Fu, “Casme i-i: An improved spontaneous micro-expression database and the baseline evaluation,” *PLOS ONE*, vol. 9, no. 1, pp. 1–8, 01 2014.

[14] Adrian K. Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap, “SAMM: A spontaneous micro-facial movement dataset,” *IEEE Trans. Affective Computing*, vol. 9, no. 1, pp. 116–129, 2018.

[15] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato, “A 3d facial expression database for facial behavior research,” in *Proceedings of Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2006, pp. 211–216.

[16] Anh Cat Le Ngo, Raphael Chung-Wei Phan, and John See, “Spontaneous subtle expression recognition: Imbalanced databases and solutions,” in *Proceedings of 12th Asian Conference on Computer Vision (ACCV)*, 2014, pp. 33–48.

[17] Shiqi Yu, Jia Wu, Shengyin Wu, and Dong Xu, “Lib face detection. https://github.com/shiqiyu/libfacedetection/,” 2016.

[18] Tomas Pfister, Xiaobai Li, and Guoying Zhao, “Recognising spontaneous facial micro-expressions,” in *Proceedings of 2011 International Conference on Computer Vision (ICCV)*, Nov 2011, pp. 1449–1456.

[19] Guoying Zhao and Matti Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, 2007.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012 (NIPS)*, 2012, pp. 1106–1114.