This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Micro-expression video clip synthesis method based on spatial-temporal statistical model and motion intensity evaluation function

PLEASE CITE THE PUBLISHED VERSION

<https://doi.org/10.1109/SMC42975.2020.9283113>

PUBLISHER

IEEE

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

LICENCE

All Rights Reserved

REPOSITORY RECORD

Wang, Lei, Jialu Hou, Xingbei Guo, Ziping Ma, Xiyao Liu, and Hui Fang. 2020. "Micro-expression Video Clip Synthesis Method Based on Spatial-temporal Statistical Model and Motion Intensity Evaluation Function". Loughborough University. <https://hdl.handle.net/2134/13387337.v1>.

Micro-expression Video Clip Synthesis Method based on Spatial-temporal Statistical Model and Motion Intensity Evaluation Function

Lei Wang, Jialu Hou, Xingbei Guo, Ziping Ma, Xiyao Liu
School of Computer Science and Engineering
Central South University
Changsha, China

{wanglei, csuhjl, guoxingbei, qdmaziping, lxyzoewx}@csu.edu.cn

Hui Fang
Department of Computer Science
Loughborough University
Loughborough, U.K.
h.fang@lboro.ac.uk

Abstract—Micro-expression (ME) recognition is an effective method to detect lies and other subtle human emotions. Machine learning-based and deep learning-based models have achieved remarkable results recently. However, these models are vulnerable to overfitting issue due to the scarcity of ME video clips. These videos are much harder to collect and annotate than normal expression video clips, thus limiting the recognition performance improvement. To address this issue, we propose a micro-expression video clip synthesis method based on spatial-temporal statistical and motion intensity evaluation in this paper. In our proposed scheme, we establish a micro-expression spatial and temporal statistical model (MSTSM) by analyzing the dynamic characteristics of micro-expressions and deploy this model to provide the rules for micro-expressions video synthesis. In addition, we design a motion intensity evaluation function (MIEF) to ensure that the intensity of facial expression in the synthesized video clips is consistent with those in real -ME. Finally, facial video clips with MEs of new subjects can be generated by deploying the MIEF together with the widely-used 3D facial morphable model and the rules provided by the MSTSM. The experimental results have demonstrated that the accuracy of micro-expression recognition can be effectively improved by adding the synthesized video clips generated by our proposed method.

Keywords—micro-expression video clip synthesis, micro-expression recognition, spatial and temporal statistical model, motion intensity evaluation, 3D facial morphable model

I. INTRODUCTION

Micro-expressions (MEs) are very subtle and involuntary facial expressions when people try to conceal genuine emotions [1]. The recognition of MEs has been widely used in many applications. For example, in the clinical field, detecting and recognizing micro-expressions is vital to assist psychologists to diagnose and remediate patients with mental diseases such as autism [2] and schizophrenia [3], [4]. ME recognition is also useful in emotion monitoring [5]–[7], teaching assistance [8], [9], serving as a vital clue in law enforcement [10], evidence collection and criminal investigations [11], [12].

Although machine learning-based (ML) [13]–[16] and deep learning-based (DL) methods [17], [18] have been widely applied for ME recognition and yielded remarkable results, these models heavily depend on the availability and effectiveness of ME datasets, which makes them vulnerable to the overfitting problem if no sufficient training data is provided. Unfortunately, it is very difficult to collect and annotate micro-expression data because micro-expressions are characterized by low intensity, fragmental Action Units (AUs), short duration and unnoticeable (only trained psychologist can perceive and decode them [19] [20], which are different from normal facial expressions). For example, One of the largest ME databases, CASMEII [21], has only 255 samples. Therefore, the scarcity of training data limits the improvement of these models for ME recognition. Transfer learning is one of the popular techniques to tackle the data shortage problem. For instance, [18], [22], [23] propose deep neural network (DNN) -based ME classifiers trained with general expression data and then fine-tuned with the limited micro-expression data. However, the effectiveness of transfer learning for ME recognition still needs improvement because the general expression domain and the micro-expression domain are substantially different in both motion intensity and the occurrence pattern of action units.

Data augmentation is another widely applied technique to address the data shortage problem. It could be classified into image-preprocessing-based augmentation and generative-model-based augmentation. Both of them still face the following difficulties when they are applied for ME recognition:

1) Methods of the former category generate new images by partially altering the appearance such as flipping images [16], adding noise [17] to increase ME images. The problem of this category is that they fail to increase the subject or emotion diversity of data, which makes the generated data lacking informative.

2) The generative-model-based methods usually synthesize facial images by tracking expressions in videos and have been widely used for synthesizing general expression data. Although this method has been used to address the data shortage problem for ME recognition, there are still three main difficulties to

This research is supported by the National Natural Science Foundation of China (61602527), Hunan Provincial Natural Science Foundation of China (2017JJ3416, 2018JJ2548, 2020JJ4746).

Fig. 1. Example frames from normal expression of happy and micro-expression of happy.

synthesize micro-expressions. (1) Micro-expressions are too subtle. As illustrated in Fig.1 (b), the difference between the onset frame and the apex frame is nearly unnoticeable, while the difference of normal expressions, such as Fig.1 (a), is apparent. It makes facial expression detection and tracking of micro-expressions much harder than normal expressions. There are no specific facial features tracking or action unit detection methods reported to be able to track or detect micro-expressions up to now. So, the generative models [24] [25] [26] which depend on facial expression tracking are difficult to capture micro-expressions and cannot produce facial images accurately. (2) The Facial Action Unit (AU) occurrence patterns of normal expressions are different from micro-expressions, which decreases the effectiveness of transfer learning from normal expressions to micro-expressions [18]. The activated AUs of the normal smile expression in Fig.1 (a) includes AU6, 12, 20, and 25, while the activated AUs of the micro-smile expression in Fig.1 (b) only includes AU12 only. (3) The intensities of micro-expressions are very small and it is difficult to synthesize facial expressions with such small intensities.

In this paper, we propose a novel micro-expression video clips synthesis scheme to address the difficulties of the generative-model-based method for ME recognition, thus overcoming the data scarcity problem. Different from the existing popular methods, our scheme does not depend on facial tracking. Instead, we propose a Spatial-temporal Statistical Model (MSTSM) to provide the rules of dynamic characteristics of micro-expressions. Applying these rules on 3D facial morphable models, we generate various facial images with micro-expressions. Furthermore, different from existing generative models, our scheme includes a novel motion intensity checking algorithm that evaluates whether the generated realistic micro-expressions which have consistent motion intensities when obeying the rules in MSTSM. The key points of our paper are as follows:

1) A novel data augmentation method for micro-expressions is proposed. This method can synthesize 3D face models and facial video clips with micro-expressions of an arbitrary number of new subjects, thereby improving ME recognition accuracy effectively.

2) To our best knowledge, this is the first ME spatial-temporal statistical model (MSTSM) to statistically analyze the dynamic characteristics of micro-expressions. The characteristics include the AUs occurrence pattern, peak intensity of each micro-expression, and the relative positions of peaks. By applying the dynamic characteristics provided by the

Fig. 2. Framework of our ME synthesis system.

MSTSM, we could produce realistic micro-expression images without using facial tracking.

3) A Motion Intensity Evaluation Function (MIEF) is designed to guarantee the accurate measurement of expression intensity in the synthesized video clips. The scheme searches in the parameters space of facial morphable models and selects the optimal parameters to generate facial expressions which is consistent with the intensity of real micro-expressions.

II. METHODS

Our scheme consists of two phases. In the first phase, we build a ME spatial-temporal statistical model (MSTSM) to represent three kinds of dynamic characteristics of micro-expressions including the action unit's occurrence pattern (AU combinations), the peak intensity of each micro-expression and the relative positions of peaks. In the second phase, we generate new ME samples with a novel motion intensity evaluation function (MIEF) and a 3D Morphable face Model (3DMM). By tweaking the shape and expression parameters of 3DMM based on the dynamic characteristics in MSTSM, facial images could be rendered via generating 3D face models with micro-expressions. Finally, MIEF is applied in the image synthesis procedure to make sure that the intensity of facial expression in the synthetic video clips is consistent with real micro-expressions. In this way, the generated micro-expression video clips with new subjects share similar dynamic characteristics with the clips of existing ME dataset. The framework of our method is shown in Fig. 2.

Fig. 3. Flowchart of MSTSM building phase

A. Build Spatial-temporal Statistical Model (MSTSM) for Micro-expressions

The ME spatial-temporal statistical model (MSTSM) is built to analyze ME dynamic characteristics statistically. We extract three types of important metrics from ME video clips, which are common AU combinations, peak intensity, and positions of apex frames. Fig. 3 shows the main steps of this phase and the detailed steps are described below.

1) Extract the Common AU combinations of micro-expressions

Facial expressions are related to the contractions of specific facial muscles, which can be described in terms of Action Units (AUs). It is noticeable that there are so many AU combinations (more than 7000) but there is no consensus about whether the AU combinations of micro-expressions are the same as the normal expressions. Therefore, we do statistical analysis to obtain common AU combinations of micro-expression in this study.

For each class of micro-expressions, such as surprise, regression, happiness, disgust, and others, we sort AU combinations by their corresponding sample number from high to low. Let $[x_1, x_2, \dots, x_n]$ denotes the sorted AU combinations and $[y_1, y_2, \dots, y_n]$ denotes the corresponding samples number, the common AU combinations $[x_1, x_2, \dots, x_m]$ are selected by:

$$\frac{y_1 + y_2 + \dots + y_m}{y_1 + y_2 + \dots + y_n} \geq \beta, \quad \frac{y_1 + y_2 + \dots + y_{m-1}}{y_1 + y_2 + \dots + y_n} < \beta \quad (1)$$

where, $m < n$, $\beta \in (0,1)$ and denotes the proportion of common AU combinations to all AU combinations.

2) Calculate the common peak intensities of micro-expressions

To quantify the intensity of an AU combination from a ME video accurately, we use optical flows as the measurement in our study. Peak intensity reflects the most intense facial movement during the process of micro-expression and thus is represented by the max optical flow intensity extracted from the aligned video clips. Following [8], all the faces in video clips are aligned

and cropped based on a reference neutral face. In this manner, the variations of spatial appearance among different video clips are normalized.

To simplify the calculation and reduce the influence of image noise, we only extract dense optical flows from the face region in each image frame where actual facial expressions may occur. Each video frame is divided into non-overlapping 8×8 blocks. The blocks that contain facial movements are called activated blocks. Only the dense optical flows in the activated blocks are computed to measure the expression intensity. For any AU combination, it is easy to tell which blocks are activated blocks by checking the regions where the corresponding AUs occurs.

The motion intensities are estimated by summing up the magnitude of optical flow vectors extracted from the first frame and each of the rest frames. The biggest motion intensity is taken as the peak ME intensity for the video clip. The intensity values are rounded to the nearest integers for ease of comparison in the later ME image synthesis phase. Finally, among the samples with common AU combinations, we count the samples number of different peak intensity and sort peak intensity by their samples number, then the common peak intensities are obtained with the similar manner as (1), except that $[x_1, x_2, \dots, x_n]$ denotes the sorted peak intensities and $[y_1, y_2, \dots, y_n]$ denotes the corresponding samples number.

3) Count the common positions of apex frames in micro-expression videos

The relative positions of apex frames is another important facial movement feature to represent the relative speeds between the contraction and expansion of facial muscles. In our work, we use a decimal to denote the position of the apex frame. Assuming a micro-expression video has L frames in total and the i -th frame is the apex frame, the position of the apex frame in this micro-expression video is computed as $R = i/L$.

For all samples with common AU combinations, we count the samples numbers of different positions of apex frame and sort the positions by their samples number, then the common

Fig. 4. Flowchart of micro-expression video synthesis with MIEF

positions of apex frames are obtained with the similar manner as (1), except that $[x_1, x_2, \dots, x_n]$ denotes the sorted positions and $[y_1, y_2, \dots, y_n]$ denotes the corresponding samples number this time.

B. Synthesize new Micro-expression video clips with 3D Morphable Model (3DMM) and MIEF

As illustrated in Fig. 4, our method include three steps to synthesize a realistic ME video clip. the first step is generating a 3D face model of an unknown subject with neutral expression by setting parameters of the 3DMM, the second step is fitting 3DMM parameters with MIEF to synthesize the 3D face model with apex expression, and the third step is rendering 2D video clip based on the generated 3D models with MEs.

1) Build the base 3D face model

Based on (3DMM) [27], the geometry and texture of a face can be represented by a shape-vector S and a texture-vector T . We define the base 3D face model with the neutral expression as (S_i, T_i) . According to the performance-based technique [28], face models (S_i^e, T_i^e) with desired expressions can be obtained by transforming the base face model linearly as in (2) and (3).

$$\Delta S_e = S_{\text{expression}} - S_{\text{neutral}}, \Delta T_e = T_{\text{expression}} - T_{\text{neutral}} \quad (2)$$

$$S_i^e = S_i + \lambda \Delta S_e, T_i^e = T_i + \lambda \Delta T_e \quad (3)$$

where λ is the parameter that controls expression intensity, the subscript “*expression*” denotes the corresponding vector related to the most intense expression, and “*neutral*” denotes the vector related to the base face model.

Micro-expression movements can be modeled by adjusting λ . When λ changes from zero to one, normal facial expressions change from neutral to apex. λ_{ME} denotes the parameter in the apex frame which has the peak intensities of micro-expressions. Because micro-expressions are much weaker than normal expressions, the parameter λ_{ME} that relates to the peak intensity of micro-expression is far less than 1. The detailed method to calculate λ_{ME} is presented in the next section.

ALGORITHM.1 3DMM PARAMETERS CALCULATION WITH EXPRESSION INTENSITY CHECKING

Input: the upper bound of λ_{ME} written as U . Initially, $U=1$
the lower bound of λ_{ME} written as L . Initially, $L=0$
the common peak intensity of a target micro-expression, written as *peak intensity* (rounded to integers)

Output: λ_{ME}

While $(U-L) \geq 0.01$
 $Median = (U+L)/2$;
 Calculate the expression intensity when $\lambda=Median$, the result saves in $Intensity_n$;
 if $Intensity_n > \text{peak intensity}$
 $U = Median$, L stays the same
 else if $Intensity_n < \text{peak intensity}$
 $L = Median$, U stays the same
 else if $Intensity_n = \text{peak intensity}$
 $\lambda_{ME} = Median$
 return λ_{ME}
 end if
end while
keep U accurate to 2 decimal places as λ_{ME} .
return λ_{ME}

2) Calculate the 3DMM parameters for the apex frames with MIEF

λ_{ME} is the 3DMM parameter relating to the peak intensities of micro-expressions. The bigger λ_{ME} is, the more intense a facial expression is and vice versa. However, there is no direct mapping between them. Here we propose an iterative searching method to find the optimal λ_{ME} , given a target ME. We name this method as Motion Intensity Evaluation Function (MIEF).

The idea of MIEF is to search for the optimal λ_{ME} which corresponds to the expression intensity equal to real MEs. The expression intensity is calculated with the optical flow between the rendered 3DMM with λ and the rendered base face model. The common peak intensities in MSTSM is regarded as expression intensity in real MEs. A binary search algorithm is applied to speed up the process. The details are described in Algorithm 1.

3) Generate Micro-expressions videos

With the calculated λ_{ME} of the apex frame, we further compute the 3DMM shape parameters for the other frames. 3DMM parameters for the other frames can be obtained based on the linear interpolation for two processes of expression movement in a video: neutral to apex and apex to neutral, using the following formulas respectively:

$$\lambda_s^{start} = \left(\frac{S}{N \times R_{ME}} \right) \times \lambda_{ME}, S = 1, 2, \dots, N \times R \quad (4)$$

where λ_s^{start} is the parameter corresponding to s -th frame in the former process, N is the total frames number in a synthetic micro-expression video, R is the position of apex frame.

$$\lambda_h^{end} = \left(1 - \frac{h}{N \times (1 - R_{ME})} \right) \times \lambda_{ME}, \quad (5)$$

$$h = 1, 2, \dots, N \times (1 - R_{ME})$$

where λ_h^{end} is the parameter corresponding to h -th frame in the process of apex to neutral.

According to the parameters sequence $\lambda_1^{start}, \lambda_2^{start}, \dots, \lambda_{N \times R}^{start}, \lambda_1^{end}, \lambda_2^{end}, \dots, \lambda_{N \times (1-R)}^{end}$, we adjust expression parameters of the 3D face model and then render the synthesized faces to image sequence: $I_1^{start}, I_2^{start}, \dots, I_{N \times R}^{start}, I_1^{end}, I_2^{end}, \dots, I_{N \times (1-R)}^{end}$ which composes the synthetic videos.

In this manner, we can synthesize ME video clips with desired face shapes, appearances, and expressions. The synthetic ME video clips have similar dynamic characteristics and similar expression intensities with real ME videos. They can help us improve the performance of ME recognition models.

III. EXPERIMENT RESULTS

Our proposed method aims to generate samples with realistic ME facial movements by matching expression intensity and dynamic characteristics with existing ME samples. The main purpose of our generated synthetic samples is to improve ME recognition performances. Therefore, we conduct experiments to evaluate if enriching ME datasets with our synthetic samples can improve ME recognition rate and exactly how much ME recognition rate can be improved in this section.

TABLE I. SPECIFIC EMOTIONS IN EACH CATEGORY AND CLIP NUMBERS IN CASME II

Category	Specific emotions
Negative (72)	Disgust (63), sadness (7), fear (2)
Others (126)	Repression (27), Others (99)
Positive (32)	Happiness (32)
Surprise (25)	Surprise (25)

A. Experimental setting

In this study, we use the CASME II dataset, one of the popular ME databases with the largest number of samples among similar databases, for ME synthesis and evaluation of micro-expression recognition improvement. The CASME II dataset consists of 255 micro-expression video clips from 26 subjects, which are recorded with a resolution of 640×480 pixels and a frame rate of 200 fps. Video clips in CASME II consists of seven classes of micro-expressions: surprise, happiness, fear, repression, disgust, sadness, and others. In our experiments, we classify CASME II into four categories: negative, positive, surprise, and others as [14]. The specific emotions that each class contains and the number of clips in CASME II are shown in Table I.

Here, we employ the methods in [14] [15], which are two of the state of the art methods, for ME recognition. Both methods extract LBP-TOP features from ME database first, then utilizes SVM to classify MEs. In our experiments, we set the proportion parameter β (defined in section II(A)) to 0.7. In total, 360 new ME samples are generated.

Section III (B) presents the statistical results of ME dynamic characteristics by using MSTSM. Section III(C) shows the comparisons between our method and the other popular augmentation methods for ME recognition.

B. The statistical result of ME dynamic characteristics

The statistical common ME dynamic characteristics are shown in Table II. In our experiments, we randomize the shape parameters of 3DMM to obtain eight new faces whose features are similar to that in the existing database. And we use Unity3D's built-in render to render the 3D models to images. Based on statistic results, we set different peak intensity for different AU combinations and only one peak intensity corresponds to one AU combination. For each new synthetic subject, we synthesized 12 (4 AU combinations $\times$ 3 position of apex frame) samples of negative, 12 (4 $\times$ 3) samples of positive, 18 (6 $\times$ 3) samples of others, 3 (1 $\times$ 3) samples of surprise.

TABLE II. COMMON DYNAMIC CHARACTERISTICS OF MICRO-EXPRESSIONS

Emotion	AU combination	Position of apex frame	Peak intensity
Positive	AU6+AU12, AU12, AUL12, AUR12	0.4, 0.5, 0.6	6524, 7425
Negative	AU9, AU4+AU9, AU4+AU10, AU4+AU7	0.4, 0.5, 0.6	5963, 5265
Surprise	AU1+AU2	0.4, 0.5, 0.6	8563
Others	AU4, AU14, AU17, AU26, AU18, AU24	0.4, 0.5, 0.6	6545, 7248

TABLE III. RECOGNITION ACCURACY OF DIFFERENT AUGMENTATIONS METHODS WITH [14] AND [15]

Augmentation method	Recognition accuracy using [14] (%)	Recognition accuracy using [15] (%)
No augmentation	65.95	46.91
Flipping [16]	65.97	47.26
Noise [17]	67.21	52.29
Proposed method	69.72	53.95

C. Comparison of different micro-expression augmentation methods for ME recognition

To evaluate the effects of our generated data for micro-expression recognition, we have tested the recognition accuracy for different expressions and generated confusion matrices.

For all ME recognition experiments, we use three-fold cross-validation which randomly divides CASME II into three folds and leave one-fold as the testing set, then combine the rest folds and synthetic samples as the training set. To the best of our knowledge, the existing methods for micro-expression data augmentation include adding noise [17] and flipping [16]. Hence, we compare ME recognition rates of our approach to these two augmentation approaches. As our method concentrates on MEs' dynamic information rather than texture information when generating new ME samples, here we employ two state-of-the-art ME recognition methods [14], [15] which pay more attention to dynamic information when recognizing MEs. The recognition results are shown in Table III.

The results in Table III show that all the above data augmentation approaches improve ME recognition performance. The best accuracy is 69.72%, which is obtained from the enrichment by our generated samples using the recognition method in [14]. The recognition accuracy of the enrichment by our generated samples is 3.77% higher than that of no augmentation. Compared with the augmentations by flipping samples [16] and noise added samples [17], the recognition accuracy of using our method is 3.75% and 2.51% higher respectively. Table III also lists the recognition accuracy of different augmentations methods of ME recognition method [15]. The recognition rate is also improved by 7.04% with our generated samples compared with no augmentation. Compared with the augmentations by flipping samples [16] and noise added samples [17], the recognition accuracy of the method [15] with our generated samples is 6.69% and 1.66% higher respectively.

These results demonstrate that our method can effectively improve the recognition performance and outperforms the other two augmentation methods for ME recognition performance improvement, which is resulted from the effectiveness of the proposed ME spatial-temporal statistical model (MSTSM) and expression intensity guarantee.

To further evaluate the effects of our generated data for micro-expression recognition, we tested the recognition accuracy for different expressions and generated confusion matrices. Fig. 5 and Fig. 6 show the confusion matrices of no augmentation, augmentation by adding noise, augmentation by image flipping, and augmentation by our method for weighted LBP-TOP [14] and LBP-IP [15] respectively.

		Prediction			
		N	S	P	O
Truth	N	50.0%	20.0%	0.0%	30.0%
	S	1.9%	57.7%	5.8%	34.6%
	P	0.0%	4.3%	91.3%	4.3%
	O	11.8%	19.1%	0.7%	68.4%

(a)

		Prediction			
		N	S	P	O
Truth	N	52.0%	4.0%	8.0%	36.0%
	S	0.0%	65.5%	5.2%	29.3%
	P	0.0%	0.0%	90.0%	10.0%
	O	13.8%	17.4%	1.4%	67.4%

(b)

		Prediction			
		N	S	P	O
Truth	N	60.9%	0.0%	8.7%	30.4%
	S	0.0%	61.5%	5.8%	32.7%
	P	0.0%	0.0%	90.0%	10.0%
	O	12.3%	21.2%	1.4%	65.1%

(c)

		Prediction			
		N	S	P	O
Truth	N	50.0%	3.3%	10.0%	36.7%
	S	0.0%	76.1%	4.3%	19.6%
	P	0.0%	0.0%	94.7%	5.3%
	O	11.6%	18.5%	1.4%	68.5%

(d)

Fig. 5. (a), (b), (c), (d) The confusion matrixes of the recognition method in [14] with no augmentation, noise-added samples, flipping samples and our generated samples respectively.

As shown in Fig. 5, compared with no augmentation, the recognition model in [14] with our generated data achieves better performances on all the four categories of MEs, which are improved by 0, 18.4%, 3.4%, 0.1% respectively. Although the recognition rates of our generated samples are lower than those of flipping and adding noise for ‘negative’ category, the enrichment by our generated samples achieves better performances on all the other three categories, which are ‘surprise’, ‘positive’, and ‘others’, by improving 10.6%, 4.7%, 1.1% for adding noise and 14.6%, 4.7%, 3.4% for flipping. These results demonstrate that the recognition method in [14] with our generated data outperforms that with no augmentation or using existing methods for micro-expression synthesis in most occasions.

As shown in Fig. 6, compared with no augmentation and the augmentation by image flipping, the enrichment by our generated samples achieves better performances on all these four categories of MEs for LBP-IP [15]. Specifically, for ‘negative’, ‘surprise’, ‘positive’, and ‘others’, the recognition rates with our generated samples are improved by 12.5%, 22.2%, 12.5%, 4.5% respectively compared with no augmentation, and 12.5%, 22.2%, 4.2%, 4.5% respectively compared with flipping. Compared with adding noise, the enrichment by our generated samples achieves better performances on ‘surprise’, which is improved by 8.1%, and comparable performances on ‘negative’, ‘positive’, and ‘others’. These results demonstrate the enrichment by our generated samples shows an outstanding ability of recognition performance improvement for LBP-IP [15], which outperforms that with no augmentation or using existing methods for micro-expression synthesis in most occasions.

		Prediction			
		N	S	P	O
Truth	N	0.0%	40.0%	0.0%	60.0%
	S	8.3%	36.1%	11.1%	44.4%
	P	0.0%	25.0%	25.0%	50.0%
	O	14.8%	24.0%	10.2%	51.0%

(a)

		Prediction			
		N	S	P	O
Truth	N	14.3%	28.6%	0.0%	57.1%
	S	9.1%	50.0%	6.8%	34.1%
	P	0.0%	40.0%	40.0%	20.0%
	O	14.6%	20.0%	10.8%	54.6%

(b)

		Prediction			
		N	S	P	O
Truth	N	0.0%	28.6%	0.0%	71.4%
	S	10.3%	35.9%	12.8%	41.0%
	P	0.0%	0.0%	33.3%	66.7%
	O	14.6%	24.5%	9.9%	51.0%

(c)

		Prediction			
		N	S	P	O
Truth	N	12.5%	25.0%	0.0%	62.5%
	S	9.3%	58.1%	7.0%	25.6%
	P	0.0%	12.5%	37.5%	50.0%
	O	14.8%	19.2%	10.4%	55.5%

(d)

Fig. 6. (a), (b), (c), (d) The confusion matrixes of the recognition method in [15] with no augmentation, noise-added samples, flipping samples and our generated samples respectively.

The experimental results show the synthetic ME video clips by our method effectively improve the performance of two of the state of the art ME recognition methods compared with no augmentation, augmentation by adding noise, augmentation by image flipping. Therefore, the effectiveness of the proposed micro-expression video clip synthesis method has been demonstrated by these experiments.

IV. CONCLUSIONS

In this paper, we propose a novel micro-expression video clips synthesis method to generate ME samples. In our method, a spatial and temporal statistical model is first established to represent the dynamic characteristics of micro-expressions such as the patterns of AUs occurrence and the peak intensities for micro-expressions. To our best knowledge, it is the first attempt to analyze and summarize the dynamic characteristics of ME samples. By using these dynamic characteristics, the 3D facial morphable model is applied to generate facial video clips to avoid using facial tracking due to this kind of tracking is difficult for micro-expressions. Furthermore, a motion intensity evaluation function is proposed to make sure that the facial expressions’ intensity in the synthetic video clips is following real micro-expressions. The experimental results have demonstrated that the generated samples by our method can effectively improve the predictive capability of the recognition model, confirming the effectiveness of our proposed micro-expression video synthesis method. For future works, we plan to improve our methods by using nonlinear interpolation to obtain the expression intensity of each frame which makes the expression intensity of synthesized samples more approach to real samples.

REFERENCES

- [1] P. Ekman, "Lie catching and micro expressions," *The philosophy of deception*, pp. 118–133, 2009.
- [2] E. Loth, L. Garrido, J. Ahmad, E. Watson, A. Duff, and B. Duchaine, "Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits?," *Molecular Autism*, vol. 9, no. 1, p. 7, Jan. 2018, doi: 10.1186/s13229-018-0187-7.
- [3] X. Zhang, L. Chen, Z. Zhong, H. Sui, and X. Shen, "The Effects of the Micro-Expression Training on Empathy in Patients with Schizophrenia," in *Man-Machine-Environment System Engineering*, vol. 456, S. Long and B. S. Dhillon, Eds. Singapore: Springer Singapore, 2018, pp. 189–194.
- [4] T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," *British Journal of Clinical Psychology*, vol. 45, no. 4, pp. 579–583, Nov. 2006, doi: 10.1348/014466505X90866.
- [5] D. M. Bernstein and E. F. Loftus, "How to tell if a particular memory is true or false," *Perspectives on Psychological Science*, vol. 4, no. 4, pp. 370–374, 2009.
- [6] G. M. A. Lodder, R. H. J. Scholte, L. Goossens, R. C. M. E. Engels, and M. Verhagen, "Loneliness and the social monitoring system: Emotion recognition and eye gaze in a real-life conversation," *British Journal of Psychology*, vol. 107, no. 1, pp. 135–153, Feb. 2016, doi: 10.1111/bjop.12131.
- [7] S. Porter and L. ten Brinke, "Reading Between the Lies: Identifying Concealed and Falsified Emotions in Universal Facial Expressions," *Psychological Science*, vol. 19, no. 5, pp. 508–514, May 2008, doi: 10.1111/j.1467-9280.2008.02116.x.
- [8] L. Dacre Pool and P. Qualter, "Improving emotional intelligence and emotional self-efficacy through a teaching intervention for university students," *Learning and Individual Differences*, vol. 22, no. 3, pp. 306–312, Jun. 2012, doi: 10.1016/j.lindif.2012.01.010.
- [9] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, Jan. 2014, doi: 10.1109/TAFFC.2014.2316163.
- [10] S. Demetriofoff, S. Porter, and A. Baker, "I know how you feel: the influence of psychopathic traits on the ability to identify micro-expressions," *Psychology, Crime & Law*, vol. 23, no. 3, pp. 274–290, Mar. 2017, doi: 10.1080/1068316X.2016.1247159.
- [11] M. Ghaziasgar, N. D. L. Cruz, A. B. Bagula, and J. Connan, "Visual Data Mining: A Great Opportunity for Criminal Investigation," *Data Mining Trends and Applications in Criminal Science and Investigations*, 2016. www.igi-global.com/chapter/visual-data-mining/157456 (accessed Apr. 27, 2020).
- [12] Qingguo, Wu, "An Analysis On Micro-Expressions in the Criminal Proceedings," *Journal of Beijing College of Politics and Law*, no. 4, p. 2, 2015.
- [13] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–6.
- [14] L. Wang, H. Xiao, S. Luo, J. Zhang, and X. Liu, "A weighted feature extraction method based on temporal accumulation of optical flow for micro-expression recognition," *Signal Processing: Image Communication*, vol. 78, pp. 246–253, 2019.
- [15] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern With Integral Projection," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Dec. 2015, pp. 1–9.
- [16] M. A. Takalkar, H. Zhang, and M. Xu, "Improving Micro-expression Recognition Accuracy Using Twofold Feature Extraction," in *International Conference on Multimedia Modeling*, 2019, pp. 652–664.
- [17] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in psychology*, vol. 8, p. 1745, 2017.
- [18] S.-J. Wang et al., "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, Oct. 2018, doi: 10.1016/j.neucom.2018.05.107.
- [19] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.
- [20] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.
- [21] W.-J. Yan et al., "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.
- [22] X. Jia, X. Ben, H. Yuan, K. Kpalma, and W. Meng, "Macro-to-micro transformation model for micro-expression recognition," *Journal of Computational Science*, vol. 25, pp. 289–297, Mar. 2018, doi: 10.1016/j.jocs.2017.03.016.
- [23] X. Ben, X. Jia, R. Yan, X. Zhang, and W. Meng, "Learning effective binary descriptors for micro-expression recognition transferred by macro-information," *Pattern Recognition Letters*, vol. 107, pp. 50–58, May 2018, doi: 10.1016/j.patrec.2017.07.010.
- [24] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.
- [25] Y. Huang and S. M. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 2259–2266.
- [26] Z. Liu, G. Song, J. Cai, T.-J. Cham, and J. Zhang, "Conditional adversarial synthesis of 3D facial action units," *Neurocomputing*, vol. 355, pp. 200–208, 2019.
- [27] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Conference on Computer Graphics & Interactive Techniques*, 1999.
- [28] F. I. Parke, "Computer generated animation of faces," in *Proceedings of the ACM annual conference-Volume 1*, 1972, pp. 451–457.