3DCNN Backed Conv-LSTM Auto Encoder for Micro Facial Expression Video Recognition

Md. Sajjatul Islam[1], Yuan Gao[2], Zhilong Ji[2], Jiancheng Lv[1], Adam Ahmed Qaid Mohammed[1], and Yongsheng Sang[1(✉)]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
sangys@scu.edu.cn
[2] TAL Education Group, Beijing 100080, China

Abstract. Facial Micro-Expression recognition in the field of emotional information processing has become an inexorable necessity for its exotic attributes. It is a non-verbal, spontaneous, and involuntary leakage of true emotion in disguise of most expressive intentional prototypical facial expressions. However, it persists only for a split-second duration and possesses fainted facial muscle movements that make the recognition task more difficult with naked eyes. Besides, there are a limited number of video samples and wide-span domain shifting among datasets. Considering these challenges, several video-based works have been done to improve the classification accuracy but still lack high accuracy. This works addresses these issues and presents an approach with a deep 3D Convolutional Residual Neural Network as a backbone followed by a Long-Short-Term-Memory auto-encoder with 2D convolutions model for automatic Spatio-temporal feature extractions, fine-tuning, and classifications from videos. Also, we have done transfer learning on three standard macro-expression datasets to reduce over-fitting. Our work has shown a significant accuracy gain with extensive experiments on composite video samples from five publicly available micro-expression benchmark datasets, CASME, CASMEII, CAS(ME)^2, SMIC, and SAMM. This outweighs the state-of-the-art accuracy. It is the first attempt to work with five datasets and rational implication of LSTM auto-encoder for micro-expression recognition.

Keywords: Micro-expression · Recognition · Deep learning · Transfer learning · Spatio-temporal

1 Introduction

Facial Micro-Expression (ME) discloses true mental state unconsciously while someone is trying to obscure them advertently in a high-stake situation. This transient ME lasts for less than 1/5 s [1] and diminishes under cover of ordinary acted facial expressions. It has a very low intensity [2] due to the tiny movements of facial muscles. It is challenging to create a high-stake situation in a controlled environment and generate ME voluntarily against its spontaneous nature. These reasons impede the real-time [3] implementation of ME recognition (MER) from publicly available ME datasets, while it gives an important

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 90–105, 2022.
https://doi.org/10.1007/978-3-031-04409-0_9

cue for lie or deceitful behavior detection. Though there are only five spontaneous publicly available ME datasets [3], all of them have smaller size of samples. Facial Action Coding System (FACS) coding [4] has also been used in most of the ME databases to identify the action units (AU) that are linked to specific facial muscle movements within facial components. AUs that are pertinent to emotional state help to reduce the subjective biases in recognition of MEs. Despite all of these challenges, ME recognition has gained momentum in the computer vision community in the last few years [5] due to some inescapable practical implications such as business deal negotiation, psychoanalysis, forensic investigation, and homeland securities.

Many handcrafted works for feature descriptions were devised based on appearance-based feature learning for ME recognition. For static and dynamic texture-based ME recognition, LBP and many of its variants have been proposed in [6–12]. Though they are very familiar, they are not innate for AU/motion-based recognition due to LBP features. Hence the geometry or motion-based descriptors [13–17] were introduced to capture the deformation in it with the facial landmarks or optical flow features. These techniques are susceptible to head-pose variations and leaned to face registration. Also, gradient-based feature descriptors [18, 19] were proposed to mitigate lighting variations but still suffer from head-pose variations.

Automatic feature learning methods ignite independent learning from inputs effectively. For that purpose, many deep learning-based models have gain popularity in recent years for ME recognition. The first possible use of the convolutional neural network (CNN) deep model [20] for ME recognition was less prone to better accuracy due to over-fitting. Takalkar and Xu [21] proposed a CNN-based model with data augmentation to combat the over-fitting problem but ended up with minor improvement due to data imbalance and subjective bias in annotations. In [22], temporal interpolation was used with DCNN followed by a support vector machine (SVM) classifier for ME recognition to combat the short duration. Peng. et al. [23] proposed a dual temporal scale convolutional neural network (DTSCNN) for Spatio-temporal feature extractions from optical flow inputs of a composite dataset from two ME datasets. It achieved better performance in comparison to some hand-design methods. In [24], CNN accompanied long-short-term-memory (LSTM) was proposed. It considered the class discrimination, expression states, and persistence of states along with temporal change. It achieved better results but was not strong enough to confront the imbalance sample problem. In TLCNN [25], pre-trained CNN was used to model spatial features from a single frame then fed them to LSTM. It used the combined video samples from three ME datasets. A pre-trained deep network-based method on apex frame was done in [26]. Two-stream optical flow-dependent high level features extractions and classification network was introduced in [27]. In [28], dual-stream shallow CNN was proposed to combat with over-fitting and saliency map. Deep recurrent-CNN (R-CNN) based-model trained from scratch for spatiotemporal feature learning were presented in [29]. In [30] shallow R-CNN model was designed and experimented on composite dataset samples. VGGNets and LSTM discriminative attention model were proposed in [31]. Deep learning-based approaches and models have shown a strong and reliable representation of discriminative features from ME.

Numerous researches have been carried out based on handcrafted feature engineering, automatic spatial features extractions, and temporal correlation among spatial features to classify the ME from SMIC [32], CASME [33], CASMEII [34], CAS(ME)^2 [35], and SAMM [36] datasets. There is some resemblance between macro-expressions (MACE) and MEs in terms of some facial dynamics. In addition, MACE datasets (e.g., CK+ [37], MUGFE [38], and OULU-CASIA [39]) have a large number of samples, which facilitate taking advantage of transfer learning in recognition of ME from limited samples using deep learning models. Automatic ME recognition from frame sequences is quite challenging due to the persistence in a small number of frames. Moreover, ME videos contain some/many redundant and neural frames that increase the computational cost as well as the influence of irrelevant feature extractions. Despite this, video-based end-to-end deep ME recognition is more resilient as it models the spatio-temporal features against the illumination, head-pose, and subtle motion variations. Though composite ME samples from multiple ME datasets make the solution even harder due to inconsistency among them, it helps to model ME to a greater extent from a diverse and large group of samples possessing different intrinsic factors that might be a set/subset in the wild samples. That paves the way to realistic categorization of ME from more spontaneous and natural ME samples.

We propose a method to extract spatio-temporal features from composite samples of five spontaneous ME datasets [32–36] through the 3D CNN in the residual network as backbone and LSTM auto-encoder for de-noising and fine-tuning the high-level feature maps comprising the temporal deformation of spatial features and followed by a native structural regularizer accompanied with a soft-max classifier. Two cross-validation strategies, Leave-One-Subject-Out (LOSO) and Stratified 5-Fold cross-validations (CV) have been designed and tested on the combined samples to estimate standard accuracy, unweighted average recall (UAR), and un-weighted average F1-Score (UF1). Two-stage transfer learning has been used based on three benchmark MACE datasets[37–39]. Our model shows superior recognition accuracy on three metrics that surpass the state-of-the-art methods.

The main contributions to this work are summarized below:

- We propose an automatic ME recognition method based on the 3DCNN-18 residual network as a backbone for spatio-temporal feature extractions followed by a conv-LSTM auto-encoder with the size of 2-1-2 for fine-tuning the temporal correlation among spatial features and de-noising them. Finally, a structural regularizer is appended for aggressive summarization to feed the final vector to a soft-max classifier.
- Macro-to-micro transfer learning has been revitalized to deal with over-fitting due to the lower number of ME video samples.
- Construction of a composite ME video dataset with three class samples (e.g., negative, positive, and surprise) from five publicly available ME datasets.
- Validation with two CV techniques, LOSO and Stratified 5-fold, to measure effectiveness and verify the generalization of the model using cross samples and cross subjects. That shows the very high effectiveness of the proposed method that outweighs the state-of-the-art MER approaches.

The rest of this paper is organized with methodology in Sect. 2, experiments in Sect. 3, and results and discussion in Sect. 4, and conclusions in Sect. 5.

2 Methodology

This section systematically presents our proposed method. The approach encompasses the pipeline, model architecture, evaluation metrics, and loss function, and transfer learning mechanism from macro to micro expressions.

2.1 Process Pipeline

General pipelines show the flow of input video processing steps, 3DCNN-18 residual [40] backbone for spatio-temporal feature extractions, 2D Conv-LSTM auto-encoder inspired by [41], and a structural regularizer followed by a soft-max classifier. It covers the input to emotional class label predictions from a sequence of spontaneous ME frames, which is depicted in Fig. 1.

Faces in ME video frames are detected and aligned with a deep-based reliable face alignment network [42] along with a real-time and robust single-shot-scale invariant face detector (S3FD) [43] to correct head-pose. It reduces the negative influence on ME recognition. The face has been cropped based on the detected bounding box to eliminate the irrelevant area from images. The facial images are then normalized to 64×64, which makes training faster for video sequences by reducing the computational cost. Then it is augmented using vertical flipping and 8° counter-clockwise rotation to triple the original samples. It helps to reduce over-fitting to some extent. Furthermore, the frame sequences have been padded with last frame which is the onset frame in our case for all ME datasets except SMIC, as the apex frame is not annotated in it. We restrict the ME video length to 12 frames for batching to reduce the computational cost and unnecessary redundancy. As the datasets samples are highly imbalanced, we have calculated the class weights for sampling in weighted random sampler for batching to ensure the representative samples for each class for recognition. After all of these preprocessing, augmentation and batching steps, the frame sequences have been fed into the model depicted in Fig. 2 for saptio-temporal feature extractions and classification. The model is described in the following section.

2.2 Model Architecture

It encompasses a residual 18 3DCNN backbone network, a 2-1-2 LSTM auto-encoder with 2D convolution, a 3DCNN layer followed by a structural regularizer, and a soft-max classifier. Here 3DCNN is for convolutions in spatial and temporal dimensions of ME frame sequence concurrently. It is implemented in a residual framework with 18 layers to capture the dynamic changes in spatial ME features from faces. The elicited spatio-temporal features with the dimension (**channel × length × height × width**) have been transferred as input to the convolutional LSTM two-layer auto-encoder for compact representation of features by reducing noises and fine-tuning the subtle spatial changes. It unrolls both LSTM encoder cells for the length dimension. We have considered the

vanilla LSTM cell with 2D convolutions, which are formulized in the following set of equations-

$$i_t = \sigma(W_{xi}\ conv\ x_t + W_{hi}\ conv\ h_{t-1} + W_{ci}\ *\ c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma\left(W_{xf}\ conv\ x_t + W_{hf}\ conv\ h_{t-1} + W_{cf}\ *\ c_{t-1} + b_f\right) \tag{2}$$

$$c_t = f_t\ *\ C_{t-1} + i_t\ *\ tanh(W_{xc}\ conv\ x_t + W_{hc}\ conv\ h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{xo}\ conv\ x_t + W_{ho}\ conv\ h_{t-1} + W_{co}\ *\ c_t + b_o) \tag{4}$$

$$h_t = o_t * \tanh(c_t) \tag{5}$$

In Eqs. (1), (2), (3), (4), and (5), it is the input, f_t is the output from forget gate, C_t current state, o_t output, and H_t is the hidden state at current time step t. Again, conv represents the convolution operation and * for Hadamard product. LSTM cell is iteratively unrolled for the activated feature map sequence obtained from the 3DCNN backbone. Conv-LSTM encoder mapped the input as a compact representation to an encoded vector. Here, it is the hidden vector h_2 from the 2nd LSTM encoder cell for the last featured frame in the high-level features frame sequence. Conv-LSTM decoder with two LSTM cells has been used to reconstruct the fine-grained, smooth spatio-temporal transient evolution. Here 3 × 3 convolutional filter is used for both the decoder and encoder parts. Regenerated ME features stacked on h_3 through **length** iterations from the decoder part are fed into a 3DCNN layer with kernel 1 × 3 × 3. It converts the feature maps from regressive to actual class form to categorize ME expression sequence into three labels negative, positive, and surprise. Then a native structural regularizer with adaptive average pooling is used for aggressive summarization to generate a vector of three predicted values. It reduces the trainable parameters to alleviate the over-fitting. A finishing soft-max classifier is used to ensure the predicted of three values is 1.

2.3 Evaluation Metrics and Loss Functions

To evaluate the model on composite ME samples, three metrics have been used to observe the influence of data organization, preprocessing, augmentation, and frame rate setup, as well as to measure the effectiveness of the proposed model. Here standard accuracy, balanced accuracy, i.e., UAR and UF1 macro metrics, have been represented in Eqs. (6), (7), and (8).

$$Standard\ Accuracy\ =\ \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$Un\text{-}weighted\ Average\ Recall\ =\ \frac{\sum per\ class\ accuracy}{C} \tag{7}$$

$$Un\text{-}weighted\ Average\ F1\ Score\ =\ \frac{2TP}{2TP + FP + FN} \tag{8}$$

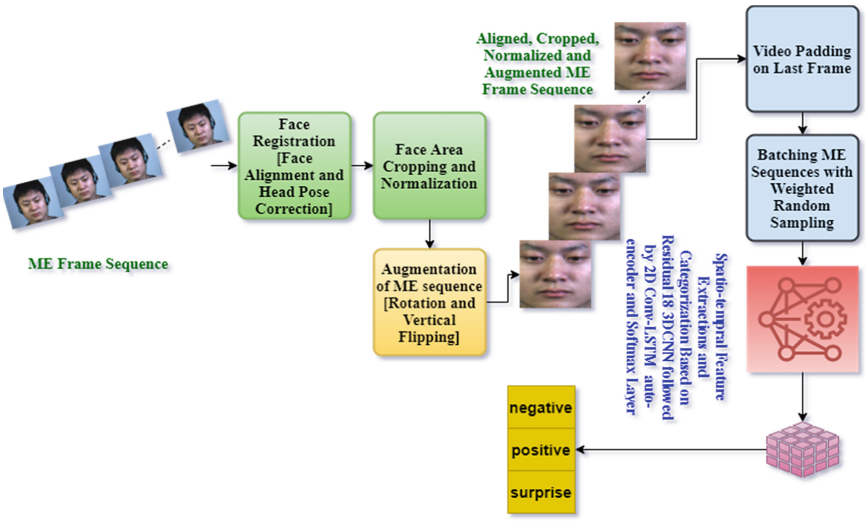


Fig. 1. ME recognition pipeline comprising preprocessing of input video to class label prediction.

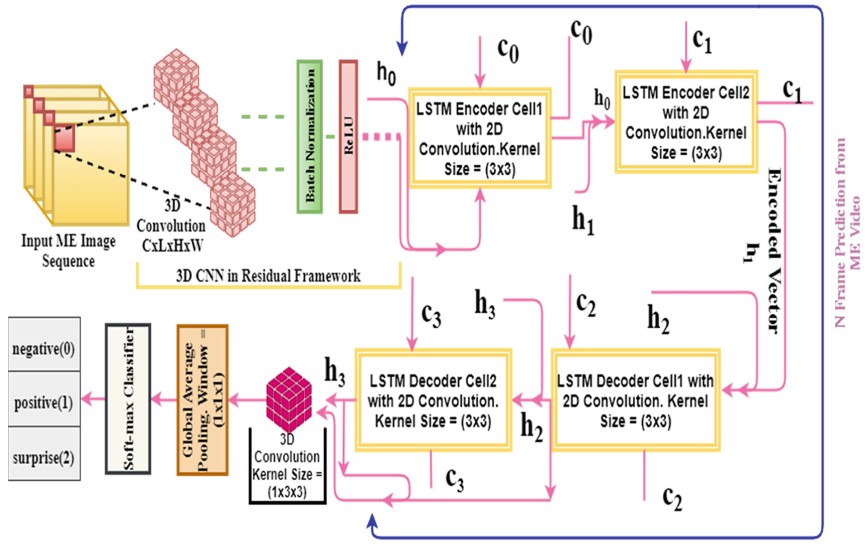


Fig. 2. Model architecture composed of 3D ResNet18 and Conv-LSTM auto-encoder appended with structural regularizer.

In Eqs. (6), (7), and (8), TP, TN, FP, FN, and C represent True Positive, True Negative, False Positive, False Negative, and the number of classes, respectively.

As the target ME datasets are highly imbalanced, we have considered these metrics to reduce the bias of the model to higher sample classes on accuracy. Model sensitivity is measured with UAR by averaging the recall on each predicted class of a ME video

sequence, and UF1 is calculated at the macro level from precision and recall for each predicted sample. Also, the standard accuracy has been estimated on the total number of correctly classified samples out of total predicted samples.

The model has been trained based on cross-entropy loss function on predicted output. Therefore, it is suitable for multiclass label classification and for highly imbalanced datasets such as publicly available ME datasets. The analytical form is given in Eq. (9).

$$loss(x, class) = -log\left(\frac{exp(x[class])}{\sum_j exp(x[j])}\right) = -x[class] + \log\left(\sum_j exp(x[j])\right)$$

For class weights-

$$loss(x, class) = weight[class]\left(-x[class] + \log\left(\sum_j exp(x[j])\right)\right) \tag{9}$$

In Eq. (9), x is the normalized output from the soft-max layer, and the class is [0, 1, 2] for three categories negative, positive, and surprise, respectively. As the model has been trained on a batch of ME frame sequence, the loss has been calculated based on the average of loss of each predicted sample in that batch.

2.4 Macro to Micro Knowledge Transfer

The model has been trained on three benchmarks MACE datasets, CK+ [37], MUGFE [38], and OULU-CASIA [39]. In our model, the backbone network is pre-trained on kinetics-400 [44], and the conv-LSTM auto-encoder with 4 LSTM cells and the 3DCNN layer are initialized randomly. The proposed model has about 37.6 million trainable parameters. On the other hand, publicly accessible ME datasets have a lower number of video samples. So, the training of the model on target ME samples in the first place must be a cause of over-fitting. Considering these facts, we have done the pre-training on MACE video samples in negative, positive, and surprise classes for similar knowledge of ME. Then the model is further trained on ME composite samples in the same categories for target knowledge MER. Datasets organization and experimental details have been discussed in Sect. 3.

3 Experiments

This section describes the organization of data samples into three discrete categories negative, positive, and surprise for both MACE and ME. It also represents the data set summaries, model implementation, and design and implementation of cross-validations LOSO and Stratified-5Fold.

3.1 Macro and Micro Datasets Organization

Three benchmark MACE datasets have been used for the transformation of knowledge about ordinary prototypical facial expressions. The widely used extended Cohn-Kanade

(CK+) [37] is one of them. The other two MUGFE [38] and OULU-CASIA [39] are also very relevant datasets for MACE video samples. We have considered frame sequences for the expressions of anger, disgust, sadness, happiness, and surprise in all three datasets. In CK+, the respective MACE videos have been reduced to the range of 6 to 50 frames by manually eliminating some early neutral frames but preserving the last one as a peak in each sequence. For MUG facial expression dataset, we have kept the video length is less than or equal to 62 frames by discarding the neutral and less expressive frames. Videos of OULU-CASIA have been restricted to 20 frames in the same way. Only the samples of the strong visible lighting part of this dataset are taken for our transfer learning. All the samples from these datasets are categorized into three classes negative (anger, disgust, and sadness), positive (happiness), and surprise (surprise). So, it results in a composite MACE three class dataset for pre-training. Samples from these representative datasets are summarized in Table 1.

We have considered five spontaneous ME datasets mentioned in Table 2 for ME recognition in negative, positive, and surprise categories. A composite dataset comprising of ME video samples from those ME datasets has been constructed. Here the negative one has been formed with anger, disgust, and sadness, where happiness and surprise for positive and surprise respectively. The number of samples in each class is tabulated in Table 1. To the best of our knowledge, this is the first attempt to recognize ME on five ME datasets. To reduce the subjective bias, all ME samples except SMIC enlisted in Table 1 have been reclassified based on AUs [45].

Table 1. Macro expressions and micro expressions video samples summary from three MACE datasets and five publicly available spontaneous ME datasets

Dataset	Negative	Positive	Surprise	Total video samples	
				Before augmentation	After augmentation
Macro expressions video samples					
CK+ [37]	133	69	83	285	1140
MUGFE [38]	376	163	143	682	2728
OULU-CASIA [39]	208	69	69	346	1384
Micro expressions video samples					
SMIC [32] (HS & VIS)	93	79	63	235	705
CAMSE [33]	80	6	14	100	300
CASMEII [34]	145	25	15	185	555
CAS(ME)^2 [35]	19	6	6	31	93
SAMM [36]	31	23	13	67	201

As ME videos have many redundant frames, which have a detrimental effect on recognition accuracy, we have eliminated many irrelevant frames by keeping only the frames in between onset and apex frames, including these two. Thus, the ME video sequences have been reduced to the intervals 6 to 13, 5 to 12, 7 to11, and 11 to 12 for CASME, CASMEII, CAS(ME)^2, and SAMM, respectively. The last frame in each ME sample is the frame with the highest peak. As SMIC is not annotated with apex frames, so we have kept the original video length. Only high-speed (HS) and visual (VIS) samples have been considered for SMIC. But this composite dataset has introduced a domain shifting challenge due to subject diversities and different disparate attributes among the ME datasets. But a larger number of samples facilitates to alleviate over-fitting and model generalization, which has been supported by the results in Sect. 4. Preprocessing and augmentation except smoothing are like those that have been discussed in Sect. 2.1. In the case of the MACE dataset, image smoothing has also been used for augmentation.

3.2 Experimental Settings

The proposed model has been trained on a composite macro dataset summarized in Table 1. The sequence length of each video is made equal to 12 frames by selecting the frames at equal interval and is padded with the last frame. Then the padded sequences have been used for pre-training. The hyper-parameters such as batch_size and initial learning rate are set to 64 and 0.0001, respectively. Learning rate is scheduled with patience 5 for equal validation loss in five consecutive epochs. An early stopping regularizing is used based on degraded or unchanged validation loss for 10 consecutive epochs. Learnable parameters have been optimized using Adam optimizer. The dimension of a batch of videos is $64 \times 3 \times 12 \times 64 \times 64$. For training and testing on ME composite target samples, we have considered the initial learning rate as 0.001, and the early stopping epoch is 8. For both cases, a weighted random sampler is configured and used to combat the bias towards the larger sample classes. But all other settings remain the same. The model is trained and tested on a platform with Windows 10 Pro, Intel Core I5 7400 CPU 3GHz, 8 GB DDR4 RAM, and NVIDIA GeForce RTX 2070 with 8 GB memory. The widely used Pythonic deep learning framework PyTorch has been used to accomplish the experiment.

3.3 Model Evaluations

The proposed model has been aligned with similar domain knowledge through the hold-out evaluations (80%:20%) on MACE composite samples. Here the model is trained and validated against three classes. The resulting confusion matrix is given in Fig. 3. Then the pre-trained model has been evaluated on our composite ME dataset in Table 1.

For this purpose, two validation strategies have been carefully considered and tuned for evaluating the model under three classes negative, positive, and surprise. One is LOSO CV, and the other one is Stratified-5 fold. Many challenges have been induced in the composite ME dataset due to the domain transition, especially the subjective diversities and the larger variations in temporal and spatial- frequencies along imbalance problem are notable. These cause a bias towards a specific group of ME video samples. LOSO and Stratified-5 fold are there to reduce biases for such data distributions. In LOSO, 87 subjects have been used to validate the model, where each subject is tested in each split during one epoch. On the other hand, each fold-out of 5 equal folds has been used as test samples in each split during one epoch. The metrics such as standard accuracy, UAR, and UF1 have been estimated from these validations. Figure 4 demonstrates the confusion metrics of best accuracies for both the tests. Also, the evolution of model convergence has been recorded in Fig. 5 and 6.

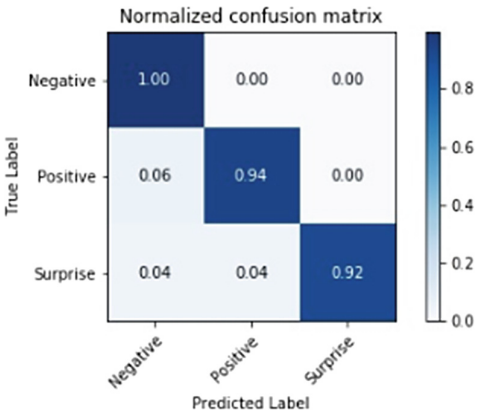
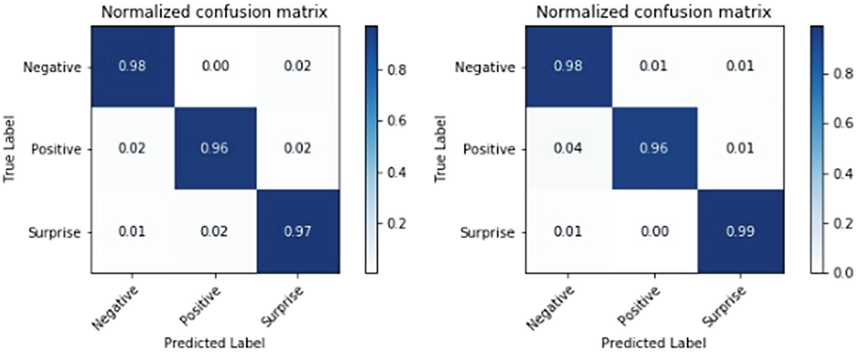


Fig. 3. Confusion matrix from pre-training on the composite MACE dataset

Table 2. Disparate spatio-temporal properties of five Publicly available spontaneous ME datasets. Here, HS-High Speed, VIS-Visual Camera, NIR- Nearly Infrared.

Spatial/Temporal properties	SMIC HS/VIS/NIR [32]	CASME [33]	CAMSEII [34]	CAS(ME)^2 [35]	SAMM [36]
Spatial resolution	640 × 480	640 × 480, 720 × 1280	640 × 480	640 × 480	2040 × 1088
Temporal frequency	100/25/25	60	200	30	200



a. LOSO CV confusion matrix b. Stratified 5-folds CV confusion matrix

Fig. 4. Confusion matrix for LOSO CV and Stratified 5-folds on composite ME dataset

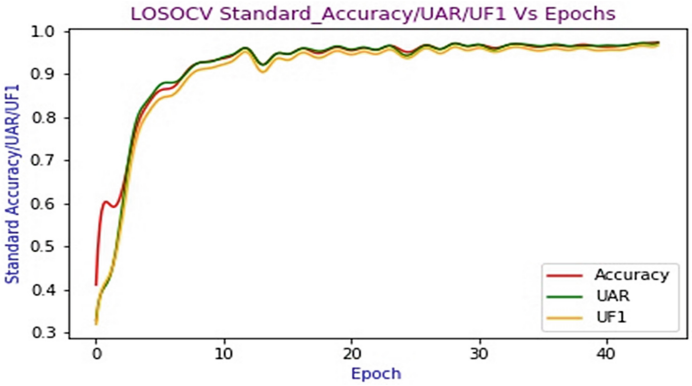


Fig. 5. LOSO CV standard accuracy, UAR, and UF1 scores

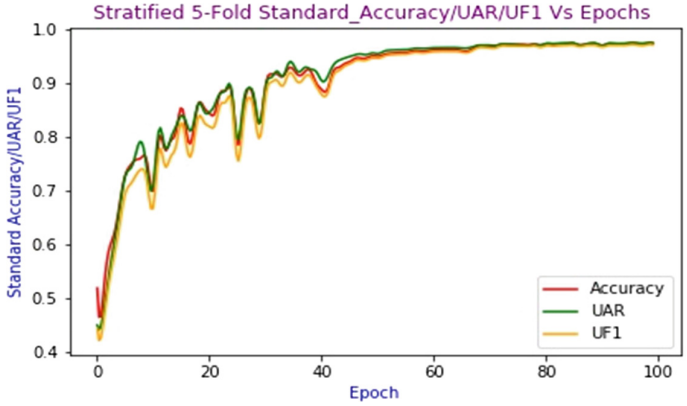


Fig. 6. Stratified 5-fold CV standard accuracy, UAR, and UF1 scores

4 Results and Discussion

We have constructed and used a composite ME dataset from the five spontaneous ME datasets mentioned in Table 3 for LOSO and Stratified 5-fold CV to categorize ME video sequence in three classes negative, positive, and surprise. Our proposed method and evaluations on this ME combined dataset show superiority to the state-of-the-art methods. It achieves significantly higher effectiveness standard accuracy, UAR, and UF1 metrics for both CVs, which is clear from the results in Table 3 and Fig. 4, 5, and 6. Table 3 has recorded the results for similar tasks base-on recent methods. CapsuleNet based on apex frame [46] has achieved about 65% UAR, and UF1 which is lower by 32% and 31%, respectively, compared to our proposed method. In RCN-A [30], the score has been improved up to 71% and 74% for both metrics. However, it is also far behind our estimated results. In macro assisted network MicroNet [47], there is a reasonable gain for both the metrics. It is ended up with scores less than 86% and 87%, respectively. Our proposed method has demonstrated a remarkable gain in terms of three metrics standard accuracy, UAR, and UF1. The estimated score is about 97% for all three metrics in the stratified CV. In the case of LOSO CV, our approach shows consistent scores of 97%, 97%, and 96%, respectively. We have covered both LOSO and stratified CV, but other approaches have done evaluations on the first one.

To lessen the over-fitting due to small size ME datasets, our composite ME dataset with larger video samples put a positive impact on all three accuracies by confronting the domain shifting challenge. Transfer learning on the integrated MACE dataset is another contributing factor to it. Subjective bias reduction with the objective categorization of ME samples, length of frame sequence reduction to 12 frames with apex frame as the last one, augmentation, and small spatial size of each frame have influenced the model prediction approvingly. With these facts, model design with spatio-temporal backbone in 3D residual framework and de-noising the five-dimensional activated feature maps with two encoders convolutional LTM layers and two accompanied LSTM decoder layers. This auto-encoder has facilitated further abstraction of relevant spatial changes in the ME frame sequence. Final structural regularization has subsidized the parameters.

The collaborative positive effects of our method on accuracies have also demonstrated the model generalization in-terms of a myriad of attributes in ME datasets, which has been captured by normalized confusion matrices and shown in Fig. 4. Some irregular weight updates have caused the minor drifting before the epoch 20 and epoch 42 in Fig. 5 and 6 respectively, which is due to the random initialization of parameters of auto-encoder and the last 3DCNN layer. After that points model is started to converge for all three metrics in both CV.

Table 3. LOSO CV ME recognition results based on contemporary methods for negative, positive and surprise classes in different compositions of ME datasets.

Methods	Composite datasets	CV	Accuracy	UAR	UF1
CapsuleNet on apex frame [46]	SMIC, CASMEII and SAMM	LOSO	–	0.651	0.652
RCN-A [30]	SMIC, CASMEII and SAMM	LOSO	–	0.719	0.743
MicroNet [47]	SMIC, CASMEII and SAMM	LOSO	–	0.857	0.864
Ours	SMIC, CAMSE, CASMEII, CAS(ME)^2 and SAMM	LOSO	**0.973**	**0.971**	**0.966**
		Stratified 5-Fold	**0.976**	**0.976**	**0.973**

5 Conclusions

Our work has proposed a method and model for ME recognition on composite ME video sequences from spontaneous datasets SMIC, CASME, CASMEII, CAS (ME)^2, and SAMM. To the best of our knowledge, it is the first attempt for ME recognition on five ME datasets. The proposed model combines the synergies of 3DResNet 18 and conv-LSTM auto-encoder. Two exhaustive cross-validations LOSO and Stratified 5-fold, have been experimented on the composite dataset. Our method shows remarkable accuracies compared to state-of-the-art methods. From our model evaluation, it is evident that the model is generalized, highly effective across domain shifting among ME datasets. In the future, this will help to classify the ME frame sequence from diverse ME samples in the wild.

Declarations.

Conflict of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This work was supported by the National Key R&D Program of China (Grant No. 2020AAA0104500), and was partially supported by Sichuan Science and Technology Major Project (Grant No. 2019ZDZX0006).

References

1. Zhang, M., Fu, Q., Chen, Y.H., Fu, X.: Emotional context influences micro-expression recognition. PLoS ONE **9**(4), 95018 (2014)
2. Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., Fu, X.: How Fast are the leaked facial expressions: the duration of micro-expressions. J. Nonverbal Behav. **37**(4), 217–230 (2013). https://doi.org/10.1007/s10919-013-0159-8

3. Takalkar, M., Xu, M., Wu, Q., Chaczko, Z.: A survey: facial micro-expression recognition. Multim. Tools Appl. **77**(15), 19301–19325 (2017). https://doi.org/10.1007/s11042-017-5317-2
4. Ekman, P., Cohn, J.F., Ambadar, Z.: Observer-based measurement of facial expression with the facial action coding system. Handbook Emot. Elicit. Assess. **1**(3), 203–221 (2007)
5. Goh, K.M., Ng, C.H., Lim, L.L., Sheikh, U.U.: Micro-expression recognition: an updated review of current trends, challenges and solutions. Vis. Comput. **36**(3), 445–468 (2020). https://doi.org/10.1007/s00371-018-1607-6
6. Pfister, T., Li, X., Zhao, G., Pietikäinen, M.: Recognising spontaneous facial micro-expressions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1456 (2011)
7. Wang, Y., See, J., Phan, R.-W., Oh, Y.-H.: LBP with six intersection points: reducing redundant information in LBP-TOP for micro-expression recognition. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9003, pp. 525–537. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16865-4_34
8. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
9. Pietikinen, G.Z.M., Huang, X., Wang, S.J.: Facial micro_expression recognition using spatiotemporal local binary pattern with integral projection. In: ICCV Workshop on Computer Vision for Affective Computing, pp. 1–9 (2015)
10. Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M.: Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns. Neurocomputing **175**(PartA), 564–578 (2015)
11. Huang, X., Wang, S.J., Liu, X., Zhao, G., Feng, X., Pietikainen, M.: Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Trans. Affect. Comput. **10**(1), 32–47 (2017)
12. Zong, Y., Huang, X., Zheng, W., Cui, Z., Zhao, G.: Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. IEEE Trans. Multimed. **20**(11), 3160–3172 (2018)
13. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, pp. 1932–1939 (2009)
14. Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X.: A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans. Affect. Comput. **7**(4), 299–310 (2016)
15. Xu, F., Zhang, J., Wang, J.Z.: Microexpression identification and categorization using a facial dynamics map. IEEE Trans. Affect. Comput. **8**(2), 254–267 (2017)
16. Liong, S.T., See, J., Wong, K.S., Phan, R.C.W.: Less is more: micro-expression recognition from video using apex frame. Signal Process. Image Commun. **62**, 82–92 (2018)
17. Happy, S.L., Routray, A.: Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Trans. Affect. Comput. **10**(3), 394–406 (2019)
18. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3D-Gradient descriptor. In: IET Seminar Digest, vol. 2009, no. 2 (2009)
19. Li, X., et al.: Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Trans. Affect. Comput. **9**(4), 563–577 (2017)
20. Patel, D., Hong, X., Zhao, G.: Selective deep features for micro-expression recognition. In: Proceedings - International Conference on Pattern Recognition, vol. 0, pp. 2258–2263 (2016)

21. Takalkar, M.A., Xu, M.: Image based facial micro-expression recognition using deep learning on small datasets. In: DICTA 2017 - 2017 International Conference on Digital Image Computing: Techniques and Applications, vol. 2017, pp. 1–7 (2017)
22. Mayya, V., Pai, R.M., Pai, M.M.M.: Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences. In: 2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, pp. 699–703 (2016)
23. Peng, M., Wang, C., Chen, T., Liu, G., Xiaolan, F.: Dual temporal scale convolutional neural network for micro-expression recognition. Front. Psychol. **8** (2017). https://doi.org/10.3389/fpsyg.2017.01745
24. Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. **10**(2), 223–236 (2017)
25. Wang, S.J., et al.: Micro-expression recognition with small sample size by transferring long-term convolutional neural network. Neurocomputing **312**, 251–262 (2018)
26. Li, Y., Huang, X., Zhao, G.: can micro-expression be recognized based on single apex frame? In: Proceedings - International Conference on Image Processing, ICIP, pp. 3094–3098 (2018)
27. Gan, Y.S., Liong, S.T., Yau, W.C., Huang, Y.C., Tan, L.K.: OFF-ApexNet on micro-expression recognition system. Signal Process. Image Commun. **74**, 129–139 (2019)
28. Khor, H.Q., See, J., Liong, S.T., Phan, R.C.W., Lin, W.: Dual-stream shallow networks for facial micro-expression recognition. In: Proceedings - International Conference on Image Processing, ICIP, vol. 2019, pp. 36–40 (2019)
29. Xia, Z., Feng, X., Hong, X., Zhao, G.: Spontaneous facial micro-expression recognition via deep convolutional network. In: 2018 8th International Conference on Image Processing Theory, Tools and Applications, IPTA 2018 – Proceedings (2019)
30. Xia, Z., Peng, W., Khor, H.Q., Feng, X., Zhao, G.: Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Trans. Image Process. **29**, 8590–8605 (2020)
31. Yang, B., Cheng, J., Yang, Y., Zhang, B., Li, J.: MERTA: micro-expression recognition with ternary attentions. Multim. Tools Appl. **80**(11), 1–16 (2019). https://doi.org/10.1007/s11042-019-07896-4
32. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikainen, M.: 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 (2013)
33. Yan, W.J., Wu, Q., Liu, Y.J., Wang, S.J., Fu, X.: CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013 (2013)
34. Yan, W.J., et al.: CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLoS One **9**(1), e86041 (2014)
35. Qu, F., Wang, S.J., Yan, W.J., Li, H., Wu, S., Fu, X.: CAS(ME)2): a database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Trans. Affect. Comput. **9**(4), 424–436 (2018)
36. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: SAMM: a spontaneous micro-facial movement dataset. IEEE Trans. Affect. Comput. **9**(1), 116–129 (2018)
37. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, pp. 94–101 (2010)
38. Papachristou, C., Aifanti, A.D.N.: The MUG facial expression database. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, pp. 1–4 (2010)

39. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikäinen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
40. Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
41. Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., Woo, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Adv. Neural Inf. Process. Syst. **2015**, 802–810 (2015)
42. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3d face alignment problem? (and a Dataset of 230,000 3D Facial Landmarks). In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2017, pp. 1021–1030 (2017)
43. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S3FD: Single Shot Scale-Invariant Face Detector. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2017, pp. 192–201 (2017)
44. W. Kay *et al.*, "The Kinetics Human Action Video Dataset," May 2017
45. Davison, A.K., Merghani, W., Yap, M.H.: Objective classes for micro-facial expression recognition. J. Imaging **4**(10), 119 (2018)
46. Van Quang, N., Chun, J., Tokuyama, T.: CapsuleNet for micro-expression recognition. In: Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019 (2019)
47. Xia, B., Wang, W., Wang, S., Chen, E.: Learning from Macro-expression: a Micro-expression Recognition Framework. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2936–2944 (2020)