# Dynamic image for micro-expression recognition on region-based framework

Trang Thanh Quynh Le
*School of Engineering*
*University of St. Thomas*
St Paul, MN, USA
le009109@stthomas.edu

Thuong-Khanh Tran
*CMVS*
*University of Oulu*
Oulu, Finland
khanh.tran@oulu.fi

Manjeet Rege
*School of Engineering*
*University of St. Thomas*
St Paul, MN, USA
rege@stthomas.edu

***Abstract*—Facial micro-expressions are involuntary facial expressions with low intensity and short duration natures in which hidden emotions can be revealed. Micro-expression analysis has been increasingly received tremendous attention and become advanced in the field of computer vision. However, it appears to be very challenging and requires resources to a greater extent to study micro-expressions. Most of the recent works have attempted to improve the spontaneous facial micro-expression recognition with sophisticated and hand-crafted feature extraction techniques. The use of deep neural networks has also been adopted to leverage this task. In this paper, we present a compact framework where a rank pooling concept called dynamic image is employed as a descriptor to extract informative features on certain regions of interests along with a convolutional neural network (CNN) deployed on elicited dynamic images to recognize micro-expressions therein. Particularly, facial motion magnification technique is applied on input sequences to enhance the magnitude of facial movements in the data. Subsequently, rank pooling is implemented to attain dynamic images. Only a fixed number of localized facial areas are extracted on the dynamic images based on observed dominant muscular changes. CNN models are fit to the final feature representation for emotion classification task. The framework is simple compared to that of other findings, yet the logic behind it justifies the effectiveness by the experimental results we achieved throughout the study. The experiment is evaluated on three state-of-the-art databases CASMEII, SMIC and SAMM.**

***Index Terms*—Micro-expression analysis, facial recognition, rank pooling, convolutional neural network.**

## I. Introduction

Facial micro-expression (ME) is one of the specific forms of facial expression. It is a very special means for people to keep their emotional states unseen. While normal macro-expression is quite evident and easy to be recognized [1], ME has been proved to be a challenge in both academic research and daily life settings [2]. The official perception of ME was first discovered and developed since 1966 by Haggard and Isaacs [3]. ME is frequently known as subtle, brief, and involuntary facial expression that occurs when people unconsciously or intentionally cover up their true feelings in high-stake situations in order to either gain advantages or avoid losses [4], [5]. Different from facial macro-expression, ME takes place in a short period of time with insignificant facial changes [4], [6]. Therefore, perceiving spontaneous ME within a particular context is a difficult task that humans have to perform. Micro-expression recognition refers to the task of identifying a ME contained in a given spotted sequence of video frames [5], and classifying the found ME into one of the predefined emotion classes [7]. Despite the aforementioned characteristics, studying ME with computational methods has increasingly acquired substantial awareness since it produces many potential values in a wide variety of applications in clinical diagnosis, business negotiation, forensic investigation, and security systems, etc. [4]. Accordingly, researchers examined tools to improve human ability of detecting and recognizing ME, yet found out that human performance would have only reached 40% with the aid of the METT training tool [8]. Consequently, the study of analyzing ME has been advanced to applying computer vision and video processing techniques, beyond psychology, to automate the process of ME analysis [6].

Deep learning has been proven to be efficient in resolving non-linear complex problems. Having emerged not long ago, it was brought into the game with the hope of becoming a promising research direction. The number of works using it to learn meaningful features has been increased aside from traditional hand-crafting feature descriptors. In [9], deep learning was first utilized in the process of learning informative features. They made an attempt to transfer pre-trained ImageNet models to extract features which later were selected by a feature selection step before feeding into a classification CNN model. The proposal obtained 56.3% in terms of accuracy rate. In addition to extracting features, many interesting works carried out deep learning as a classifier to tackle the problem of automatic ME recognition. [6] considered the representation of a ME by its single apex frame. The attention was brought back to a single apex frame where rich features could be learned exclusively and fed into the fine-tuned deep VGG-Face architecture. This method attained 63.3% accuracy. Another solution was introduced that adopted CNN models to learn small scale spatio-temporal feature representations with their expression-states taken into account. Spacial features and their states were encoded by a CNN model and later transferred to learn temporal characteristics. The learned features were subsequently fed into long short-term memory recurrent neural

*The first two listed authors contributed equally to this work.

networks to classify and achieved a classification rate of 60.98% [9]. Other works have also substantially contributed to automatic micro-expression recognition from different perspectives using different fine-tuned neural networks, yet the results obtained were not satisfactory [7]. The poor performances of the proposed frameworks might be caused by the subtlety and low intensity of ME [6] and insufficient small-sized databases [7].

In this paper, we propose a succinct yet efficient end-to-end framework for automatic facial ME recognition as illustrated in figure 1. By taking advantage of dynamic image to summarizing video sequences, we simplify the ME recognition task and turn it to a single image classification problem. In this fashion, several state-of-the-art deep learning models can be facilitated to smoothly recognize ME samples. The framework has proved to be effective as we achieved an impressive recognition rate 78.5% on CASMEII, even though it is very easy to implement compared to other complicated frameworks with hand-crafted feature descriptors. The result will be discussed in depth in subsequent sections.

The remaining sections of this paper will be constructed as follows: Section 2 reviews the related works; Section 3 introduces the methods used for our proposed framework; Section 4 presents the experimental results and discussions, and Section 5 is the Conclusions.

## II. Related Work

### A. Region-based approach

Characterized as subtle and brief motion, ME is proved to be region-dependent [10], because its occurrences are fast, involuntary [5], and rarely take place in the entire face. Many research have been finding a way out to extracting expensive features on local segments as opposed to a holistic facial representation. Therefore, [11] proposed that facial micro-expression analysis should be operated on upper and lower halves of a face independently in lieu of dealing with the entire face in order to get good results. According to the work of [5], the face has been divided up to a specific number of blocks. However, there are certain facial areas that do not contribute much to the ME analysis [7], such as cheeks. To address this dilemma, [12] found that the face should be decomposed into a few region-of-interests (ROIs), and the ROIs that contain facial motions triggered when one or more facial expression coding system (FACS) action units are activated should come into the picture [7]. ROIs has been shown to be more efficient in ME spotting against the entire face [12], [13]. Analyzing certain ROIs is predominantly applied to spotting task because it helps accurately detect geometrical features [14]. However, only a limited amount of research took advantage of local patches in the task of ME recognition such as Zhao and Yu using necessary morphological patches [2]. In [1], Wang et al. used ROIs in close cooperation with micro-attention mechanism for ME recognition task and obtained a 66% accuracy rate. We also find ROIs powerful in helping our framework effectively discriminate ME. Hence, we adopt this approach to put our concentration on a few facial patches in stead of the whole face.

### B. Feature extraction methods

In automatic ME recognition task, finding a good feature extractor is always considered an integral part. Many studies have taken advantage of hand-crafted feature extraction methods. In particular, LBP-based and optical flow-based are the two mostly applied methods. Local binary pattern (LBP) [15], an appearance-based method, extracts local texture changes out of circular regions with binary codes being encoded into a histogram. Several variants of LBP were derived, but the most popular method which was also presented as the baseline evaluation method for a big number of works is LBP-TOP [16]. It is the extended version of LBP built upon the spatio-temporal domain over three orthogonal planes. This allows LBP-TOP to dynamically exploit temporal constraints along with spatio neighborhoods. On the other hand, optical flow-based methods measure intensity variation, relying heavily on temporal dynamics in video sequences [7]. Other methods are also frequently employed for feature representation such as Histogram of image gradient orientation [5]. Those feature descriptors deploy the spatio-temporal characteristics of a video sequence to extract local texture and temporal dynamics.

Previous works also explored those feature representations in their frameworks to advance the ME recognition. Pfister et al. [17] applied LBP on three orthogonal planes (LBP-TOP) to extract features, followed by a multiple kernel learning algorithm for a binary classification. Three widely-used feature descriptors, namely LBP-TOP, HOG and HIGO-TOP were chosen for the experiments in Li X. et al. (2018) [5]. In addition to traditional feature extraction methods, a few research also employed deep learning in learning hidden features. [9] transferred the pre-trained ImageNet models for feature learning. In our work, a deep learning-oriented technique is applied to effectively learn long term dynamics of a video sequence.

### C. Deep learning architectures

The rapid growth of deep learning and CNN in a course of time have been undeniably and incredibly beyond imagination. Deep learning has been promoted and developed in various applications including ME analysis, and become one of the popular methods for face recognition in particular. There has been a number of deep learning architectures introduced in computer vision arena such as VGG-Net [18], ResNet [19], Senet50 [20]. Although those networks approached quite differently with respect to architectural design and computational proposition, they were all inherent the ability and the capability of learning spacial features in hidden layers that lead to high end results for image applications. ME recognition research has also progressively contacted CNN either to acquire a better feature extraction method or to accomplish a better recognition rate as a classifier. CNN, in the first place, was introduced in the work of [9] to extract informative features from sequences of frame. Later, Peng et al. [21] advanced the use of

CNN with an end-to-end network called Dual Temporal Scale Convolutional Neural Network for ME recognition task. Deep learning has really been accelerated to a different level and outperformed other methods in computer vision domain. With ME recognition task particularly, it is expected to surpass the limitation of insufficient databases and exceed the boundaries set by challenging tasks proposed.

## III. Proposed framework

In this section, we introduce the details of our framework for the ME recognition problem. The proposed method includes four main components: (1) Facial Motion magnification to highlight the motion in ME samples. (2) Rank pooling to compute the dynamic image which is the main feature in our method. (3) We employ the Region of Interests (ROIs) to extract the specific regions from the facial parts. (4) Convolutional neural network (CNN) architectures are constructed to recognize emotion types. In our framework, we utilize the face aligned data from original datasets; therefore, we do not need to apply the face preprocessing steps in our framework.

In the next subsections, we discuss each component in details.

### A. Facial motion magnification

The low intensity characteristic of ME makes it hard to be observed. To address this problem, Eulerian video magnification method [22] is used to enlarge small facial movements in input sequences. The idea is implemented with the combination of both spatial and temporal processing. For any given input video sequence, spatial decomposition is carried out to obtain different segments namely spatial frequency bands which might exhibit different ratios. A bandpass filter is applied on those spatial bands to extract frequency bands of interest considering time series of intensity values in each pixel. The extracted spacial bands are then amplified by a given magnification factor $\alpha$ before added back to the original signal and collapsed to generate a final output.

Motion magnification has been proved to significantly boost ME recognition performance. This Eulerian magnification method is widely used as one of the fundamental preprocessing steps because it tremendously helps magnify subtle motions by enabling the differences between distinct categories to be more effectively discerned. We adopt it for video magnification to enlarge subtle ME contained in each sequence as to address the problem of low intensity, and thus increase ME recognition rate.

### B. Dynamic image for feature extraction

Feature extraction and classification are two dominant players that have a big impact on ME recognition task. Most of the current feature descriptors, as discussed in section 2, were developed based on spatio-temporal domain as temporal template reasoning about temporal changes while spatial aspect measures intensity changes. The mostly applied descriptor, LBP-TOP, supports this concept and dynamically encodes temporal variations in video sequences. However, the underlying mechanism of these traditional feature extraction techniques only capture local changes in small time windows. They have the capability to learn dynamic textures but seemingly fail to capture long-term motion patterns corresponding to any particular action. To this end, we employ the dynamic image in [23] for our framework in order to extract the gist of video sequences as we are convinced that temporal variations and long-term dynamics are essentially equally distributed in terms of performance contribution.

Dynamic image is a simple and deep learning-oriented feature representation based on rank pooling method. It is constructed from the image pixel level stacking up RGB channels of each pixel in a frame and later projecting all elements of the frames onto a larger vector d*. This is a real vector and contains the same number of components as the frames. Because of that, vector d* is capable of ranking all frames in the sequence yet still maintaining the dynamics of the whole video. The process of assembling vector d* is called rank pooling, and d* is used to extract content and appearance information in frame sequences.

In this proposed system, dynamic image is used to pool information from magnified image sequences. Particularly, a sequence of frames is fed into the descriptor for feature capture after being magnified. By aggregating spatial features and projecting RGB components of each frame in a sequence over a large vector, the descriptor now becomes a compressed version of the whole sequence and is represented as a standard RGB image.

### C. Region of interests extraction

Due to the subtle characteristic of ME, it rarely takes place in the entire face. Researchers have proposed to localize face area and analyze ME based on facial patches in contrast to one entity. It is observed that particular emotions usually get triggered and associated with specific subsets of facial muscles corresponded to facial action units (AUs). In that way, any particular ME is defined to get involved with a subset of AUs such that when those AUs are triggered within the regions of interests, a facial motion corresponding to that ME occurs.

ME occurs in the first place when a person consciously or accidentally hides their true feelings. Hence, ME is produced with low intensity as much as possible. Because of that objective, only a few facial regions are involved in the process of producing ME. We are strongly convinced that not all information on the face is important, and with region-based approach, we can place the concentration on the facial locations that significantly contribute to the feature learning stage, and reduce unnecessary motions generated by uninteresting areas that might affect the whole performance of the task.

Therefore, in this framework, we apply region of interests to only extract specific regions for classification. In our study, targeted region are forehead, eyebrows, nose, mouth and mouth corner areas as illustrated in figure 2. Eyes are recognized to be the most noticeable and salient facial region than its counterparts for automatic facial micro-expression spotting.

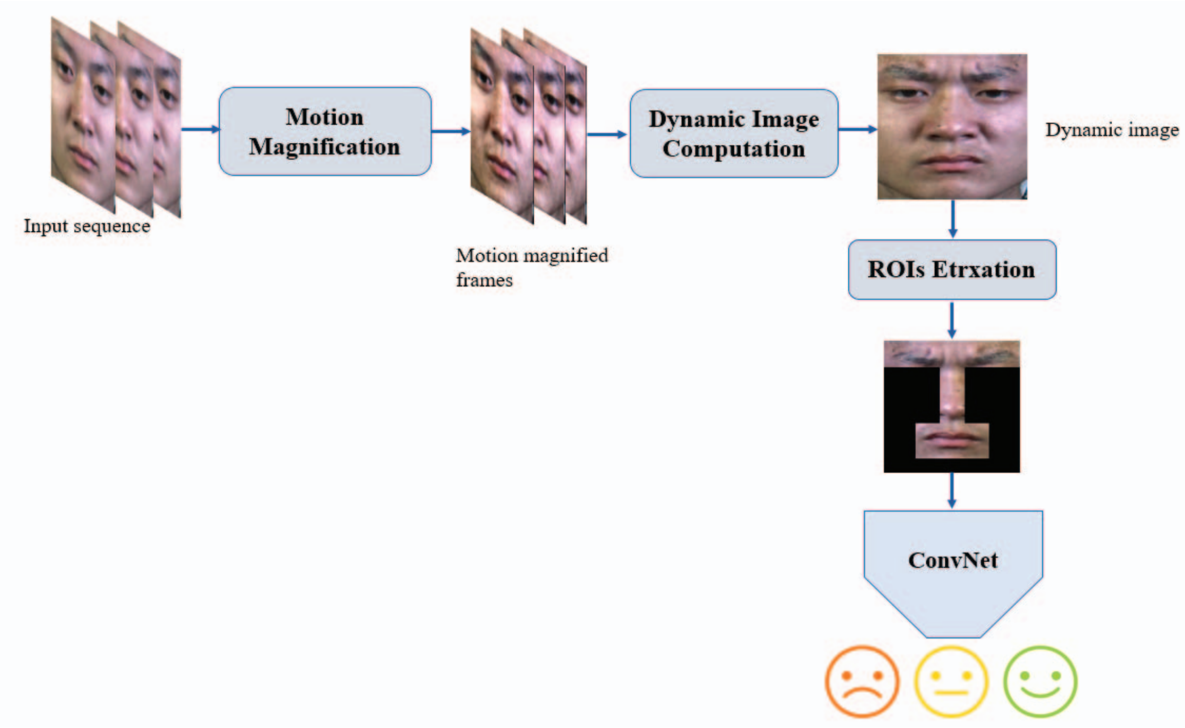

Fig. 1. Illustration of four main components in the proposed method for ME recognition system. MEs in an input video sequence are first magnified using Eulerian video magnification, which helps enlarge facial movements for the succeeding feature learning phase; The output frames then go through computations for extracting features based on rank pooling called dynamic image technique. The outcome of this stage is RGB dynamic images representing features of each frame in the sequence; Certain localized facial regions are next extracted on the elicited dynamic images: forehead, eyebrows, nose, and mouth for highlighting dominant facial motions existing in the frames. A CNN model is run on the pre-processed data for emotion categorization as a final stage.

However, eyes are masked in our study since eye cascades and blinking are not ME and usually cause false alarms.

In Fig. 2, three Region of Interests (ROIs) are defined from the landmark points on the face. In practice, we utilize DLIB toolbox to extract 68 landmark points. Then, we use the points corresponded to eyebrow, noise and mouth for extracting the information.

### D. Classification

With rank pooling operated as a feature descriptor, the outcome for each video is a single dynamic RGB image comprising the gist of the video. Traditional machine learning algorithms might not completely be capable of taking full advantage of those motion compressed images. They might work well with static images since they behave as a classifier on straightforward feature vectors. However, when dynamic images get involved, they could not capture long term dynamics therein and lose the merit of that respect. Moreover, rank pooling method introduces non-linearity into the framework, so conventional algorithms like random forest, logistic regression, etc. would not be efficient.

We leverage the advancement of deep neural networks for recognizing and classifying ME. Specifically, CNN acts as a subordinate feature learner and a classifier. Convolutional layers in CNN can learn pixel intensities corresponding to ranking hierarchy extracted in previous step. Therefore, the elicited dynamic image after being patched for region of interests will be going through another learning phase and classifying phase afterward for categorizing ME into predefined emotion classes. In the experiments, three advanced architectures for image classification are used to construct the model: VGG19, ResNet50 and Senet50.

## IV. Experiments

In this section, we discuss in details the implementation as well as results of the experiments. This section is structured as two parts: (1) the experimental settings and implementation details are presented clearly in the first part. (2) The experimental results are explained in the second half.

### A. Experimental settings and implementation

*1) Settings:* Having a sufficient and representative ME database is half way through solving any problems in an automatic ME recognition task. A number of well-labeled spontaneous ME databases have been publicly introduced for academic purpose. However, they are still relatively limited and small-scaled compared to resources used for solving other computer vision problems. There are several publicly available databases that are considered benchmarks for ME analysis performance evaluation. For our study, we select three state-of-the-art spontaneous ME databases CASMEII [24], SMIC [17]

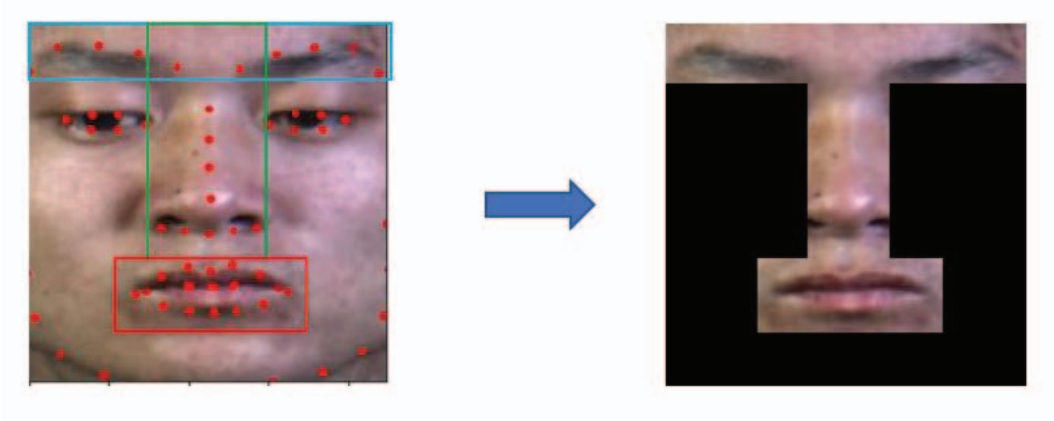

Fig. 2. Illustration of the Region of Interest component, which is used to apply in the dynamic image. Following landmark points, three regions corresponded to eyebrow (blue rectangle), noise (green rectangle) and mouth (red rectangle) are utilized to crop the ROIs for next training.

| Databases | CASMEII | SMIC-E-HS | SAMM |
|---|---|---|---|
| Sample | 247 | 164 | 159 |
| Subject | 26 | 16 | 32 |
| Emotion | 5 | 3 | 7 |
| FPS | 200 | 100 | 200 |
| FACS Coded | Yes | No | Yes |

TABLE I
CASMEII, SMIC AND SAMM COMPARISON

and SAMM [14] for the experimental evaluations. CASMEII is the largest spontaneous ME database up to date with 247 videos and 26 participants, which contains both AU labels and emotion labels. Videos were recorded with high quality and at 200 fps per sample. The number of emotion classes is 5 classes: happiness, disgust, surprise, repression and other. SMIC, developed in 2017, comprises up to 164 MEs from 16 subjects of different racial backgrounds. The database was extended and consisted of 3 versions HS, VIS, and NIR with each version recorded at different FPS. The emotion labels were negative, positive and surprise. In this paper, we evaluate our framework on the HS version. SAMM, on the other hand, is considered the most culturally diverse database that includes 159 videos from 32 subjects. While all selected participants came from a homogeneous ethnic background in CASMEII, subjects in SAMM were from 13 different ethnicities. The videos were recorded at the same 200 fps for each, with a high spatial resolution. The emotions were then labeled by a trained professional as anger, contempt, fear, disgust, happiness, sadness, surprise, and others. Table 1 compares the 2 databases in more details.

Inspired from the previous studies [5], [25], we regroup emotion types having similar behaviors into one common group. Particularly, there are 4 classes in CASMEII as opposed to 5 original classes. We unify disgust and repression into a single negative class. SAMM is marked with 7 emotions and get combined to become 4 class labels as disgust, anger, sadness, and fear are merged into negative class. Contempt are grouped into other, while surprise and happiness remain unchanged. For effectively evaluating our cross-domain experiments with SAMM and CASMEII, we regroup MEs into 4 categories: positive, negative, surprise, and others similarly as [21], [25], [26]. With respect to SMIC dataset, the authors categorized emotions into three sub-groups: positive, negative and surprise. Hence, we proceed the experiment without further relabelling.

The experiments are carried out on CASMEII, SMIC and SAMM exclusively. Only the results of CASMEII and SMIC are presented and analyzed with other existing works that approached the same evaluation method. To show an unbiased settlement, we also conduct cross-database experiments with SAMM and CASMEII each in turn being training and testing sets, then compare them with the one that use equivalent evaluation protocol. This valuating manner provides a broader and objective way to properly understand and estimate performances of the framework.

*2) Implementation:* We follow the evaluation protocol from the study [25] to conduct our experiment and analyze the results. In this protocol, 90% of the samples are assigned to training set, and the remaining are used for testing. K-fold cross-validation (K=10) is employed to avoid biases in method evaluation. Additionally, we conduct cross-database evaluation to explore performance of our proposed method in various domain.

To extract ROIs of dynamic images, DLIB toolbox is used to predict 68 landmark points on the face. Subsequently, facial regions of forehead, eyebrows, nose and mouth area are extracted by the specific points. In training deep learning architectures, Pytorch framework is utilized to build CNN models. We use Tesla T4 GPU to train our model. The models are rigorously trained with 30 epochs. As to accelerate the training process, different optimizers are tried out, and stochastic gradient descent is finally selected with learning rate 0.0001.

### B. Experimental result

To assess our method, we compare our results to reported results from previous studies as in [21], [25]. Relevant comparisons are shown in Table II.

With respect to experimental results on CASMEII, we realize that our method produces impressive outcomes. VGG19 is the best model obtaining the first place accuracy ($78.5\%$). The runner up is Senet50 with $67.3\%$ accuracy while in previous studies, Dual Temporal CNN model achieved $66.67\%$ accuracy and was considered the top performer. On SAMM dataset, VGG19 is still leading with the accuracy of $61.20\%$, closely followed by Senet50 model. There is no existing study evaluating their method by the same protocol, we only report our result.

For performance evaluation on SMIC, we find that our methods are also better than the existing works [25]. On Table II, the model VGG19 also obtains the first place by accuracy $72.69\%$. On the second place, Resnet50 is located with performance $69.49\%$. Compared with the study [25], accuracy is $70.5\%$, our proposed technique is more effective.

We proceed experiment without ROIs and analyze the performance of our framework. The achieved accuracy rate is far from being as good as that involving ROIs; thus, we do not report result of that experiment. We notice that using the entire face in extracted dynamic image will lead to noises caused by unnecessary movements from irrelevant facial areas and consequently weaken the classifier. As a result, we conclude that having ROIs defined and extracted in the framework is essential.

Following the results, we come to the conclusion that our approach is better than the compared methods. It is obvious that VGG surpasses other architectures regarding facial analysis topics. In previous research, it was widely utilized for facial expression recognition. VGG is by nature terrific at discriminating sophisticated hidden information in data. Therefore, along with motion magnification and fast dynamic image computation, small movements of facial ME can be effectively distinguished and discerned.

Experimental results of our method across CASMEII and SAMM for cross-database evaluation are displayed in Table III. Each column name implies the order we use for training and testing respectively. The first dataset is the training set, and later is for testing. Compared to the existing study [26], our technique yields better results when exploiting ResNet50 architecture trained on SAMM. Our finest model (VGG19) obtains the best accuracy by $53.75\%$ compared to $42.35\%$ of that in [26]. However, when swapping training and test sets, the performance slightly descends ($52.45\%$) versus $53.46\%$ in their study.

## V. Conclusion

We present a compact framework that yields impressive results for ME recognition system. By complying the concept of dynamic image containing motion information in image sequence with state-of-the-art deep neural network, our framework rigorously exhibits a robust and effective yet easy-to-implement process analyzing and classifying ME. In the experiments, the reported results show that our method outperforms the existing studies.

However, there is still room for future improvements. Despite the significant outcome the proposed framework produces, the faraway satisfaction is not yet anywhere adjacent to our grasp. We need to explore a way to end the gap and accelerate results to closely align with the reality. Furthermore, we are fully aware of the advantage as well as the challenge of using ROIs. Yet we arbitrarily apply the self-defined template to manually obtain facial regions; thus the accuracy might be slightly affected. We suggest combining AUs detection knowledge in our system to improve the classification performance.

## References

[1] C. Wang, M. Peng, T. Bi, and T. Chen, "Micro-attention for micro-expression recognition," *arXiv preprint arXiv:1811.02360*, 2018.

[2] Y. Zhao and J. Xu, "An improved micro-expression recognition method based on necessary morphological patches," *Symmetry*, vol. 11, no. 4, p. 497, 2019.

[3] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of research in psychotherapy*. Springer, 1966, pp. 154–165.

[4] P. Ekman, "Lie catching and microexpressions," *The philosophy of deception*, vol. 1, no. 2, p. 5, 2009.

[5] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 563–577, 2017.

[6] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?" in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3094–3098.

[7] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: Databases, methods, and challenges," *Frontiers in psychology*, vol. 9, p. 1128, 2018.

[8] P. Seidenstat and F. X. Splane, *Protecting airline passengers in the age of terrorism*. ABC-CLIO, 2009.

[9] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2258–2263.

[10] P. Husák, J. Cech, and J. Matas, "Spotting facial micro-expressions "in the wild"," in *22nd Computer Vision Winter Workshop (Retz)*, 2017.

[11] 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, 2008.

[12] A. Davison, W. Merghani, C. Lansley, C.-C. Ng, and M. H. Yap, "Objective micro-facial movement detection using facs-based regions and baseline evaluation," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 642–649.

[13] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Automatic micro-expression recognition from long video using a single spotted apex," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 345–360.

[14] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2016.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[16] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[17] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *2011 international conference on computer vision*. IEEE, 2011, pp. 1449–1456.

TABLE II
COMPARISON BETWEEN OUR PROPOSED METHOD AND SEVERAL EXISTING WORKS IN THREE DATASETS.

| **Method** | **CASMEII** | **SMIC** | **SAMM** |
|---|---|---|---|
| STCLQP with codebook [27] | 64.78% | 64.02% | - |
| FHOFO with SVM [28] | 55.86% | 51.22% | - |
| DT-CNN [21] | 66.67 % | 53.6% | - |
| Facial Color + LSTM [25] | 66.66 % | 70.5% | - |
| Resnet50 + ROI | 59.45% | 69.49% | 55.45% |
| Senet50 + ROI | 67.3% | 62.25% | 57.2% |
| VGG19 + ROI | **78.5%** | **72.69%** | **61.20%** |

TABLE III
COMPARISON BETWEEN OUR PROPOSED METHOD AND EXISTING WORK WHEN UTILIZING CROSS DATABASE EVALUATION.

| **Method** | **SAMM/ CASMEII** | **CASMEII/SAMM** |
|---|---|---|
| 3DCNN [26] | 42.35% | **53.46**% |
| ResNet50 + ROI | 47.5% | 51.25% |
| VGG19 + ROI | **53.75%** | 52.45% |

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[21] 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.

[22] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, “Eulerian video magnification for revealing subtle changes in the world,” *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[23] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic image networks for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034–3042.

[24] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, “Casme ii: An improved spontaneous micro-expression database and the baseline evaluation,” *PloS one*, vol. 9, no. 1, 2014.

[25] H. Shahar and H. Hel-Or, “Micro expression classification using facial color and deep learning methods,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[26] Y. Wang, H. Ma, X. Xing, and Z. Pan, “Eulerian motion based 3dcnn architecture for facial micro-expression recognition,” in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 266–277.

[27] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, “Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns,” *Neurocomputing*, vol. 175, pp. 564–578, 2016.

[28] S. Happy and A. Routray, “Fuzzy histogram of optical flow orientations for micro-expression recognition,” *IEEE Transactions on Affective Computing*, 2017.