**International Journal of Multimedia Data Engineering and Management**
Volume 11 • Issue 4 • October-December 2020

# Rank-Pooling-Based Features on Localized Regions for Automatic Micro-Expression Recognition

Trang Thanh Quynh Le, University of St. Thomas, USA
Thuong-Khanh Tran, University of Oulu, Finland
Manjeet Rege, University of St. Thomas, USA

## ABSTRACT

Facial micro-expression is a subtle and involuntary facial expression that exhibits short duration and low intensity where hidden feelings can be disclosed. The field of micro-expression analysis has been receiving substantial awareness due to its potential values in a wide variety of practical applications. A number of studies have proposed sophisticated hand-crafted feature representations in order to leverage the task of automatic micro-expression recognition. This paper employs a dynamic image computation method for feature extraction so that features can be learned on certain localized facial regions along with deep convolutional networks to identify micro-expressions presented in the extracted dynamic images. The proposed framework is simple as opposed to other existing frameworks which used complex hand-crafted feature descriptors. For performance evaluation, the framework is tested on three publicly available databases, as well as on the integrated database in which individual databases are merged into a data pool. Impressive results from the series of experimental work show that the technique is promising in recognizing micro-expressions.

## KEYWORDS

Dynamic Image, Image Classification, Micro-Expression

## 1. INTRODUCTION

Facial micro-expression (ME) is a momentary facial movement that delicately conveys emotions. If recognizing normal macro-expressions is relatively effortless (Wang, Peng, Bi et al, 2020) as they are apparent and noticeably obvious, it is seemingly tricky to identify different MEs within certain contexts (Zhao & Xu, 2019). The first ME occurrence was acknowledged in 1966, when discovered by Haggard and Isaacs in a filmed interview (Haggard & Isaacs, 1966). It shortly became well known and well formed in psychology. ME occurs in the first place when a person either unconsciously or intentionally hides their genuine feelings as to obtain some personal goals or avoid dangers (Ekman, 2009),(Li et al., 2017). Contrary to macro-expressions, MEs arise over a transient moment with meager facial muscular changes (Ekman, 2009),(Li et al., 2018). Hence, discerning spontaneous ME in any particular situation is a difficult and complicated activity.

DOI: 10.4018/IJMDEM.2020100102

ME analysis task is fundamentally partitioned into spotting and recognition. After an ME being spotted from an input video sequence, it is detected (Li et al., 2017) and categorized into several predefined emotion labels (Oh et al., 2018). Regardless of the striking attributes and the complexity of analyzing ME, researchers have been extensively studying ME using computational methods, as it is promising and applicable in many disciplines such as security systems, clinical diagnosis, forensic investigation, etc. (Ekman, 2009). In order to aid humans in distinguishing different MEs, various tools have been invented of which performances were not sufficient. The best training tool was METT with 40% of MEs being correctly recognized (Seidenstat & Splane, 2009). As a result, studying ME has been leveled up to employ cutting-edge techniques in computer vision to automate the task and achieve better outcomes.

Deep learning has emerged and rapidly become a rising approach for solving challenging problems. With high-level feature representation, deep neural networks have leveraged a lot of automatic tasks including micro-expression recognition. (Patel et al., 2016) was the first paper that made an attempt to apply deep learning in learning rich features. They utilized transfer learning by taking pre-trained Image-net models, and applied it to feature extraction. After that, a feature selection step was implemented to select only relevant information, before feeding them all into a convolutional network for classification. The proposed framework achieved a 56.3% accuracy rate. Aside from feature learning, many studies have engaged deep learning with ME recognition as a classifier. In (Li et al., 2018), a VGG-Face model was run on features exclusively learned from single apex frames out of input sequences.They assumed that apex frames contained the most important facial information, and by neglecting other unnecessary frames they could avoid supplemental noises leading to a better outcome. This method obtained a better detection rate, 63.3%. Another work adopted different CNN architectures to extract low-level features based on spatio-temporal domain taking into consideration their expression states. Spatial features were encoded along with their expression states by a CNN architecture and afterward used to learn its corresponding temporal texture. A LSTM network was subsequently employed to classify emotions and gained an outcome of 60.98% accuracy. Other research have also significantly endeavored to tackle and improve the task of automatic micro-expression recognition by using different approaches and fine-tuned networks. Nevertheless, the results were not sufficiently satisfactory (Oh et al., 2018). The undeniably tough characteristics of ME, low intensity and short duration (Li et al., 2018), in addition to the lack of ME data (Oh et al., 2018) account for the impairment in performance of previous methods.

In this paper, we introduce a straightforward, simple and compact yet very effective framework for automatic facial ME recognition, as illustrated in the Figure 1. By applying the rank-pooling based dynamic image to extract the gist of video sequences, we simplify the task of identifying ME. With the amount of steps involved reduced, we turn it into a single RGB image classification problem. After feature extraction, elicited images are then further learned by descriptor in the form of convolutional layers in posterior deep convolutional networks. They classify MEs at the end as a classifier. The implementation is simple compared to other intricate methods with low-level feature extraction techniques.

Characterized as being subtle and transient, ME data is difficult to be naturally induced and simulated. Not having enough data, a lot of methods poorly performed in the task of ME recognition even though proposed feature descriptors were capable of obtaining informative feature vectors. Additionally, two out of three state-of-the-art databases SMIC and CASMEII, used for evaluation in this study, were produced with constraint in racial background as well as participant maturity. That would lead to the models being biased in making inaccurate predictions when confronting subjects from unfamiliar ethnicities and ages. For those reasons, an integrated database is generated with the intention of solving the limitations caused by the lack of ME data. Data from SMIC, CASMEII and SAMM are combined in such a way that can neglect their distinct frequencies. After computing dynamic images from individual databases, they are merged together to classify MEs.

The remainder is organized as follows: Section 2 discusses some existing works related to the methods employed in proposed framework; Section 3 introduces each step and talks about the reasoning behind them; Implementational setup and experimental results are discussed in section 4; and conclusion is drawn in the last part.

## 2. RELATED WORK

### 2.1. Feature Extraction Techniques

Expressive information is learned on input face images and interpreted as features when a model classifies MEs. Thus, having a good feature descriptor significantly determines the efficiency and renders an ME recognition framework well-performed on the task at hand. Most of the feature extraction methods have derived and developed based on spatio-temporal domain. With LBP-based and optical flow-based families being the two frequently applied methods, they have predominantly been exploited for ME recognition. While optical flow-based methods describe the intensity variation while embracing temporal dynamics (Oh et al., 2018), local binary pattern (LBP) (Ojala et al., 2002) counts on appearance textures which uses binary pattern to present local facial shift out of circular areas then encodes those extracted patterns into a final histogram. A big number of LBP-based variants were evolved. Of all the derived versions, local binary pattern over three orthogonal planes (LBP-TOP) (Zhao & Pietikainen, 2007) became a baseline evaluation technique for many studies in the field. It exploits the spatio-temporal domain on three orthogonal planes: spatial plane (XY), vertical spatio-temporal plane (YT) and horizontal spatio-temporal plane(XT), allowing temporal variations to be encoded dynamically.

Those low-level feature representations focus on facial shift on pixel level. They grasp the spatio-temporal aspect of video sequences to extract features, hence are mainly deployed in ME frameworks to take advantage of tiny conceivable facial motions existing in frame sequence. In particular, LBP-TOP was applied as a feature descriptor in (Pfister et al., 2011) followed by a multiple kernel learning method for classification. Li X. et al. (Li et al., 2017) exclusively selected LBP-TOP, HOG and HIGO-TOP for their experiments. Aside from conventional low-level feature representations, some other works also adopted deep learning into their frameworks. (Patel et al., 2016) used transfer learning to learn informative features from pre-trained ImageNet models. Deep learning and CNN rapidly became popular in the field as they help accelerate the learning process and dramatically boost up accuracy rate.

### 2.2. Facial Localization Approach

ME occurrences are involuntary (Li et al., 2017), transient and rarely take place in the entire face, it is perceived as region-dependent (Husák et al., 2017). As a result, many studies have turned to facial localization instead of holistic facial representation. (Porter & Ten Brinke, 2008) suggested that ME analysis can be executed independently on either upper or lower parts as opposed to the entire face. According to (Li et al., 2017), the whole face was partitioned into smaller blocks. Yet, it was argued that some portions did not strongly impact ME analysis (Oh et al., 2018). To this end, (Davison et al., 2018) proposed that the entire face should be disintegrated into region of interests (ROIs). The ROIs that get triggered when one or more facial action units get activated should be discriminated and utilized to extract rich features (Oh et al., 2018). In the task of ME spotting (Davison et al., 2018) (Liong et al., 2016), ROIs have effectively demonstrated its efficacy in discriminating facial movements compared to the whole face in spotting task. Adopting ROIs helps eliminate dispensable local regions and reduce noises that might have been added to feature vectors otherwise. Although it is prevalent and common in spotting task, the number of research that took advantage of ROIs in ME recognition is yet bounded. Zhao and Yu used necessary morphological patches (Zhao & Xu, 2019) to extract local patches in their framework. Wang et al. employed ROIs in conjunction

with micro attention mechanism (Wang, Peng, Bi et al, 2020) and obtained an accuracy of 66% for classification. We find ROIs powerful and helpful in highlighting salient facial local patches so we apply this approach in the proposed framework.

### 2.3. Deep Learning Networks

The emergence of deep learning urged researchers to look for some optimal ways in order to extract features and advance the task in the field of computer vision. Having been promoted and developed in various practical applications, deep learning in general and convolutional neural networks (CNN) in particular have intensively become a popular method for image identification. Subsequently, a variety of neural networks have been proposed and thrived such as VGG-Net (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), etc. which made a huge contribution and changed the way researchers approach tasks in the arena. Although those networks were designed and implemented differently in terms of computational composition, they retain the ability to capture high-level features and capability to take use of multiple hidden layers for better feature learning and representations. As a consequence, previous studies have also leveraged ME recognition with the use of deep neural networks. Particularly, CNN was employed in the first place for feature extraction in the study of (Patel et al., 2016). Peng et al. (Peng et al., 2017) invented an end-to-end dual temporal scale CNN to accelerate feature learning for a better high end result. Deep learning has leveled up the task of ME recognition and boosted up the performances of other works. With the lack of ME data and their special attributes, deep learning has been an aspiration to surpass the inherent boundaries.

## 3. PROPOSED FRAMEWORK

In this section, we present the reasoning behind our motivation of choosing specific techniques laid out in the framework. There are four main components: (1) Facial motion magnification to magnify small facial changes in video sequence; (2) Feature extraction using the dynamic image computation; (3) Region of interests (ROIs) are applied to select distinct motions as final feature vectors for CNN; and (4) different CNN architectures are executed for classification task.

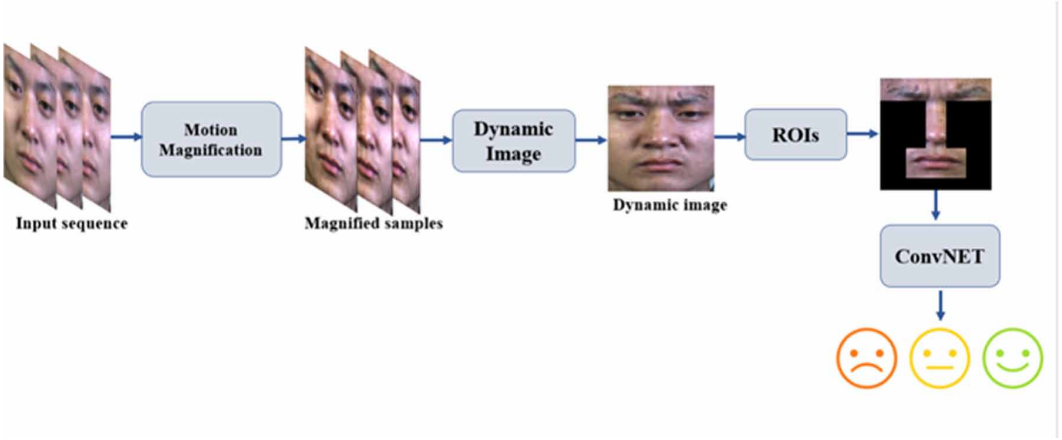
### 3.1. Facial Motion Magnification Method

ME takes place over a short period of time with insignificant facial changes. The low intensity characteristic makes it hard to perceive spontaneous ME within a particular context. To address this problem, Eulerian video magnification (EVM) method is used to enhance miniature facial movements in input sequences. This method exploits both spatial and temporal facets to process videos. Spatial decomposition is enforced to acquire different spatial frequency bands which possess distinct ratios. Those spatial frequency bands are later filtered out by a bandpass to obtain intriguing frequency bands while taking into account intensity variation over temporal values at pixel level. A magnification factor $\alpha$ of choice is used to amplify the extracted frequency bands prior to merging them back to the original vector.This method significantly helps magnify subtle facial movements and enable distinct motions to be more effectively observed. Therefore, it is often used as one of the fundamental preprocessing steps in ME recognition frameworks. Characterized as being subtle, ME can be really hard to be discerned by naked eye. With EVM, we find it practical and effective in rendering ME more distinguishable. Hence, in this framework, we adopt it for motion magnification as to address one of the prominent characteristics that differentiates ME from normal facial expression, low intensity.

### 3.2. Feature Extraction with Dynamic Image

In emotion recognition and ME recognition in particular, the ability to extract good features leading to correctly identifying facial motions plays an important role in ascertaining how well the framework performs. It has been devoted by a massive pool of studies attempting to finding optimal ways

**Figure 1. A sequence of frames or video is going through motion magnification step using Eulerian video magnification for movement magnification. The outcome is returned images with enlarged facial movements. Subsequently, the dynamic image method is employed to extract feature textures from the magnified frames/videos. The intermediate outcome of this stage is a single RGB dynamic image containing all information of the input frames. The extracted features are later filtered out considering only four targeted local facial regions (ROIs). It generates a final feature vector which only retains information ultimately fed into ConvNet models for emotion categorization.**




to depict texture features in videos. The majority of current well-known feature descriptors were developed based on spatio-temporal domain. While spatial facet grasps intensity changes in pixels, temporal aspect pays attention to variations with respect to time. LBP-TOP, the largely exploited descriptor, upholds this mechanism and accomplishes considerably positive results. However, these traditional feature extraction methods can only capture pixel variations within small time intervals. Their underlying mechanism allows them to pick up local dynamic textures yet fails to learn long-term patterns, in which the essence of the whole video associating with major motions in video sequences is showcased. To solve this problem, as we are convinced that long-term dynamics are indispensable and significantly contributing to increasing overall performance, a technique called dynamic image is applied, which extracts features on a broader and holistic level.

Dynamic image is a computation that simply summarizes video content by remodeling video structure to be a typical static image which represents all information of the whole video. This feature resembling is basically based on the idea of a ranking function proposed in (Fernando et al., 2015), where each frame in a given input video is ranked and ultimately mapped to a real vector that carries all appearance and distinctive elements from them. Specifically, from T frames in a video, T feature vectors are extracted and averaged over time t. The ranking function pairs each time t with a score $S(t \mid d)\}$ such that it demonstrates hierarchy of the frames. For later time t+1, corresponding score $S(t+1 \mid d)+$ will be greater than the preceding one. d, in this context, is a vector containing all parameters that can be learned to correctly manifests the ranking. It is learned as an optimization problem with a regularizer term and a hinge loss.

$$E\left(d\right) = \frac{\lambda}{2} \parallel d \parallel^2 + \frac{2}{T\left(T-1\right)} \times \sum max\{0, 1 - S(t+1 \mid d) + S(t \mid d)\} \tag{1}$$

The objective is to ensure that interval t+1 has a better score than time t by a margin of 1. It is then mapped to a single vector $d^*$ :

$$d^* = argmin_d E\left(d\right) \tag{2}$$

Vector $d^*$ is a real vector which carries all information and distinctive elements from the input frames and compresses them into a single shared depository. Hence, it consists of all content and components of the individual frames. Due to that procedure, vector $d^*$ is allowed to rank all the frames in a sequence yet still retain the dynamics in it. The course of constructing vector $d^*$ is known as rank pooling, and is used to extract features in frame sequences. In this framework, as magnified frames are fed into the descriptor, dynamic image is employed to pool information from them. By piling up spatial aspect over temporal factor, the feature vector becomes a compact visual representation embodied in the form of a standard RGB image.

### 3.3. Localized Facial Region Extraction

ME happens when a person accidentally or consciously masks their true emotions. It is produced with initial intention of low intensity, and rarely occupied by the entire face. Therefore, only a few facial areas are engaged in ME delivery and get triggered for particular emotions. We believe that not all information on the whole face is equally important. Few subsets of facial muscles are associated with certain facial movements and corresponded to a number of facial action units (AUs). Accordingly, when AUs associated with specific facial regions get triggered, the correspondent ME takes place.

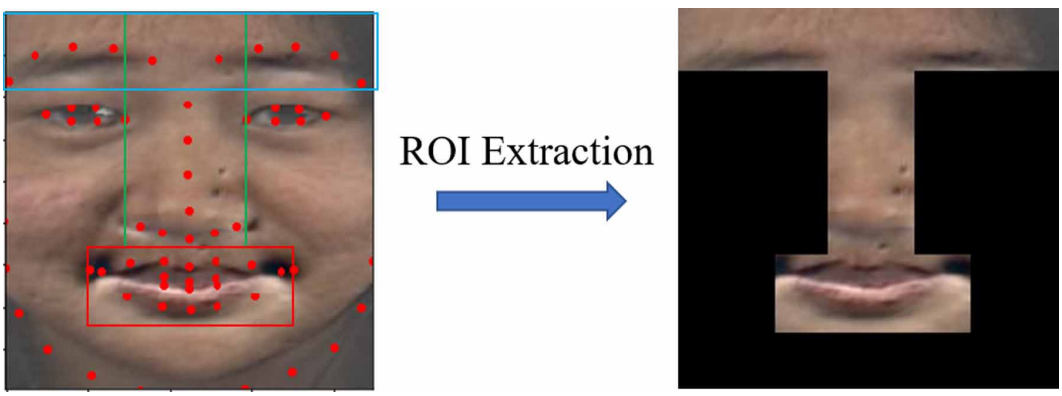
On that account, researchers have approached ME recognition task with localized facial patches as opposed to one single face entity. With region-based method, we can concentrate on facial parts that substantially contribute to motion generation leading to performance improvement. Consequently, we employ region of interests (ROIs) to extract intriguing facial locations before the classification. In this framework, the targeted regions are forehead, eyebrows, nose and mouth, as illustrated in Figure 2. Even though eyes are considered the most noticeable part out of the entire face, we decide to mask them off. Due to eye cascades and blinking, they would introduce false alarms into the framework resulting in potential performance descending.

For implementation, we use DLIB toolbox to extract 68 facial landmark coordinates. After that, only specific landmarks related to targeted ROIs are taken into account for information extraction. In Figure 2, three distinct ROIs are outlined from facial landmarks.

### 3.4. Emotion Classification

With dynamic image learning temporal information from the frames, extracted features comprise the gist of the video, which is the dynamics of facial motions. As the concept of dynamics gets involved,

**Figure 2. Extracting facial motions in four regions: forehead, eyebrows, nose and mouth**



traditional machine learning (ML) methods might not be efficient. They might perform well on still images as the features are straightforward, but seemingly fail to capture long-term patterns in local motion based features. In addition, conventional ML algorithms can learn to reasonably predict the frames on an individual level. However, they are supposedly incapable of connecting the dots between consecutive frames, which is crucial in ME recognition. As ME is described as being very subtle, typically there is insignificant to slightly no change of facial movements in multiple frames in a video sequence. For that reason, the final prediction which is either averaged out or the mostly voted outcome would possibly be inaccurate. Moreover, dynamic image would aggregate non-linearity into ME data, resulting in CNN being a promising candidate to resolve this.

During the training process, CNN serves as a subordinate feature descriptor after dynamic image and also as a final classifier. In particular, convolutional layers can automatically learn pixel variations corresponding to ranking hierarchy pooled in the last step. Hence, input of this stage is the dynamic image being patched by four attentive ROIs. The elicited image will be extracted for a richer feature representation which is eventually going through a final phase for categorization. For experiment implementation, five advanced deep networks are applied VGG19, ResNet50, ResNet34, SeNet50 (Hu et al., 2018) and EfficientNet(Tan & Le, 2019).

## 4. EXPERIMENTS

Details of experimental implementation and results are discussed in this section. The first part will be explaining the settings and implementation. The results will be stated clearly in the second half.

### 4.1. Setting and Implementation

In deep learning applications, data is a critical criterion in producing good results. In ME recognition, several small to medium-sized ME databases, which are both posed and spontaneous, have been introduced and utilized in various studies. However, when an ME is deliberately simulated, it somehow loses the natural flow no matter how well the subjects demonstrate it, and the intensity will not be as low as that of spontaneous ME. The networks running on posed data will unexpectedly pick up different traits corresponding to features learned from input frames. Accordingly, only spontaneously produced ME data is utilized in our experiments.

A number of well-known ME databases are publicly introduced for ME analysis and used as benchmarks for evaluating a number of ME work. Characterized as being brief and subtle, naturally producing and capturing ME is very challenging. Those databases are very limited in quantity compared to other large visual datasets available in image recognition. In this study, we select three state-of-the-art ME databases for framework evaluation: CASMEII (Yan et al., 2014), SMIC (Li et al., 2013; Pfister et al., 2011), and SAMM (Davison et al., 2016).

SMIC is one of the standardized ME databases widely used for performance evaluation in a variety of studies. It was conducted in 2017 with 3 versions: high speed (HS), near-infrared (NIS) and normal visual (VIS). In this paper, we evaluate the proposed framework with HS subset. Even though they were recorded with different FPS, they shared the same resolution. The data were labeled as positive, negative and surprise, yet did not provide AUs annotation.

CASMEII is the biggest ME database in size with 247 videos conducted by 26 subjects. It was well labeled with both emotion labels and AUs annotation. The videos were captured with high speed (200fps) and high quality. There are 5 defined emotions, namely happiness, surprise, disgust, repression, and other, in which some classes were not fairly distributed among others. SAMM is the culturally diverse database in which 32 participants came from heterogeneous racial backgrounds. The database consisted of 159 ME samples recorded at 200 FPS with high resolution. They were then labeled as anger, contempt, fear, disgust, happiness, sadness, surprise, and others. Details of the selected databases are displayed in Table 1.

**Table 1. ME database overview**

| | SMIC (HS) | CASMEII | SAMM |
|---|---|---|---|
| **Subjects** | 16 | 26 | 32 |
| **Samples** | 164 | 247 | 159 |
| **Labels** | 3 | 5 | 7 |
| **AUs** | No | Yes | Yes |

Similarly in (Li et al., 2017) and (Shahar & Hel-Or, 2019), we regroup emotion labels that possess comparable natures and gather them into one common class. Specifically, in CASMEII, we combine disgust and repression into one collective category, negative, reducing the number of labels from 5 originally to 4 classes. For SAMM, we merge disgust, anger, sadness and fear into one negative class as they express unpleasant sentiments, lessening labels from 7 down to 4. On the other hand, we group contempt into other, whereas surprise and happiness stay unchanged. With SMIC, there are only 3 basic emotion categorizations: positive, negative and surprise. Hence, we do not need any further re-labelling on this database.

In order to justify the unbiased settlement, we carry out a cross-database experiment on CASMEII and SAMM. For convincingly validating the results, emotions in both of the selected databases are regrouped into 4 classes: positive, negative, surprise and others as inspired from (Peng et al., 2017), (Shahar & Hel-Or, 2019), and (Wang, Ma, Xing et al, 2020). The experiment is done on SAMM and CASMEII, each being training and testing set and vice versa, and comparing the results with existing research that had the equivalent evaluation protocol on this comparable analysis. This experiment reveals a point of view where we can understand more thoroughly whether or not the performance of the proposed framework is heavily dependable upon trained data.

Deep learning participates in the framework as a notable piece. It is obvious that data play an essential part in making any deep learning application robust. Yet, ME data is notoriously known as complicated in inducing and labeling. Additionally, because of the characteristics of ME, parameters such as high resolution and high frame per second (FPS) are required to meet some standards. Those hurdles remarkably contribute to the limitation of rendering ME analysis frameworks. In consequence, we decide to merge all data into one data source and proceed the evaluation on the mixed database. Specifically, three mentioned databases are integrated into a big blended pool of data that CNN models can benefit from. Due to the fact that CASMEII, SMIC, and SAMM were all recorded with different FPS, CASMEII and SAMM were recorded 200 while SMIC's FPS is 100, dynamic images are extracted separately for each database before merging into the bulk. They are independent of one another despite the heterogeneity since at this point, the focus turns to temporal information and summarized local motions of the frames instead of the difference of frame parameters.

We follow the evaluation protocol from the study (Shahar & Hel-Or, 2019) to conduct our experiment and analyze the results. In this protocol, 90% of the samples are assigned to the training set, and the remaining are used for testing. In addition, K-fold cross-validation (K is set to 10) is employed to avoid biases in method evaluation, and cross-database evaluation is conducted to explore performance in various domains.

In training deep learning architectures, Pytorch framework is utilized to build CNN architectures. We utilize Tesla P100 GPU to train our models. The models are rigorously trained with 200 epochs. As to accelerate the training process, different optimizers are tried out, and stochastic gradient descent is finally selected with learning rate 0.001.

## 4.2. Experimental Results

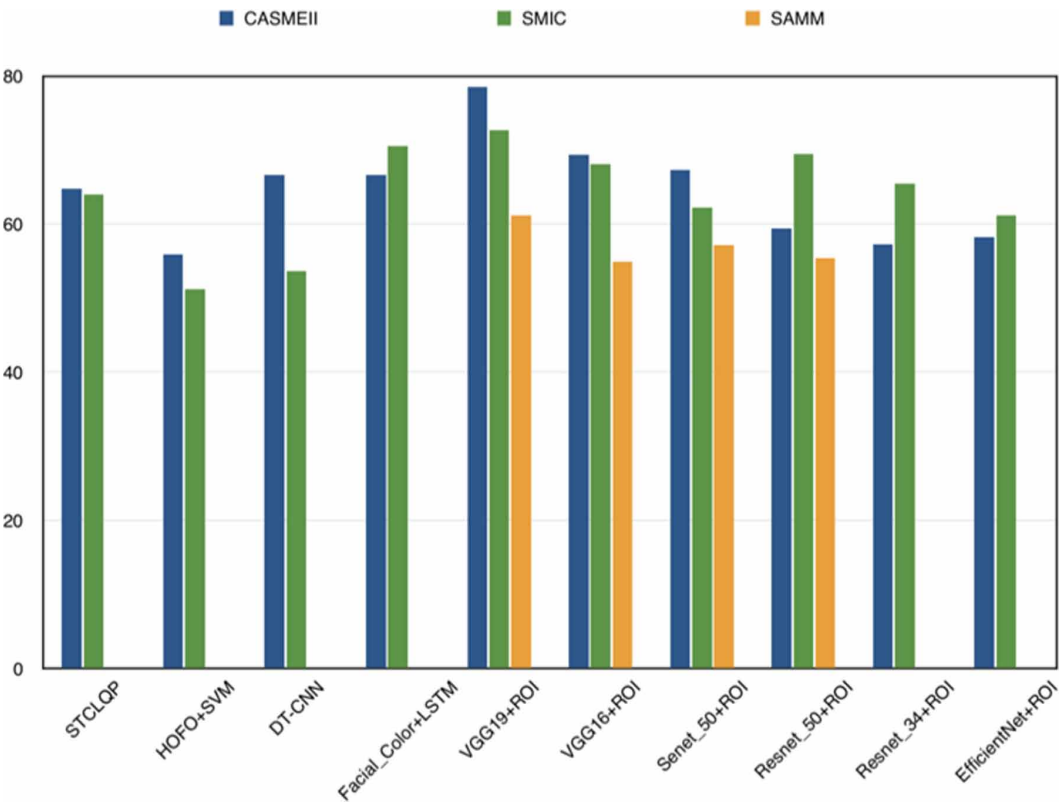
In this section, performance results of our proposed method are recorded and assessed based on the rate differences from other methods.

### *4.2.1. Stand-Alone Database Results*

In order to analyze the proposed framework, some existing works that share the same evaluation protocol are chosen for performance analogy. Selected works are STCLQP with codebook (Huang et al., 2016), HOFO with SVM from (Happy & Routray, 2017), DT-CNN in (Peng et al., 2017), and Facial Color + LSTM in (Shahar & Hel-Or, 2019).

As demonstrated in Figure 3, our method shows excellent results and remarkably surpasses other compared studies. Specifically, it achieves an impressive result on CASMEII and wins the first place with VGG19 obtaining 78.5% accuracy rate. VGG16 follows up with 69.3%. Both of those results outperform the remaining ones including the proposed method in (Peng et al., 2017) holding the best rate (66.67%). VGG19 keeps leading the board on SMIC with 72.69% accuracy, which is better than the succeeding method proposed in (Shahar & Hel-Or, 2019) viewed as the top head (70.5%). The second best goes to Resnet50 with 69.49%. With respect to SAMM, perhaps due to the nature of data which is of gray scale, it might not be the database of choice for a variety of ME works. Additionally, there is little to no existing work that has compatible assessing rules as ours, we only report attained results from our framework. VGG19 gains a relatively pleasing result (61.20%), closely followed by Senet50 with 57.2%.

**Figure 3. Illustration of our performances along with other existing studies**



#### *4.2.2. Cross Database Evaluation*

Indicated previously, we also evaluate the proposed method on cross-database protocol. In Table 2, the analogy between our method and the other is displayed. CASMEII and SAMM are selected with each successively being training and testing sets. The order in column names defines the role of each dataset in the training process.

It can be seen that with SAMM operated as training data and CASMEII being testing data, all CNN architectures deployed on our method exceed the compared method (Wang, Ma, Xing et al, 2020). VGG19 discloses the best accuracy of 53.75%, apart from the result in (Wang, Ma, Xing et al, 2020) by a large margin (>11%). However, with training and testing data displacement, the performance inconsiderably decreases and becomes less as good.

**Table 2. Result comparison in cross database evaluation**

| Method | SAMM / CASMEII | CASMEII / SAMM |
|---|---|---|
| **3D-CNN (Wang, Ma, Xing et al, 2020)** | 42.35% | **53.46%** |
| **VGG19 + ROI** | **53.75%** | 52.45% |
| **VGG16 + ROI** | 51.45% | 51.85% |
| **Resnet_50 + ROI** | 47.5% | 51.25% |
| **Efficient + ROI** | 48.24% | 51.9% |

#### *4.2.3. Integrated Database Evaluation*

In Table 3, the evaluation is presented on the mixed database. It is seen that models fed with features patched by ROIs doubtlessly display better outcomes in comparison with the ones that do not have ROIs engaged. Among the top performers with ROIs, ResNet50 returns the best outcome with an impressive accuracy rate (60.81%). Along with that, VGG19 produces a good result as well with 58.87% accuracy.

### 4.3. Discussion

As to understand the effect of ROIs on our framework, we proceed with experiments where no ROIs are involved. Features are directly fed into CNN models for classification after the extraction of dynamic image. The results obtained in individual databases are not solid, so we do not show the details. Nevertheless, we have the accuracies described for the integrated database in Table 3, even though they appear to be low compared to the models that have ROIs. When extracting features on a holistic perspective, all facial information is processed despite the weight of each feature. Since ME is notoriously subtle, it is better represented with local motion based methods. ROIs help tune out trivial contributions of features on facial regions that do not participate in ME production and also reduce noises caused by unnecessary movements from irrelevant facial areas consequently lead to weakening the classifier. Therefore, having ROIs defined is essential in improving the classification rate.

With the experimental results laying out, we have the pleasure to conclude that our framework outperforms other existing studies. VGG19 successfully exhibits great accuracy rates on all three databases and cross database as well. Regarding the integrated database performance, Resnet50 generated the highest rate. Having deep layers stacked on top of each other, these two backbones are well known in solving face recognition and become image classification benchmarks.

**Table 3. Comparison of our proposed method in combined database evaluation**

| Method | Accuracy |
|---|---|
| **VGG19** | 42.25% |
| **Senet_50** | 48.56% |
| **Resnet_50** | 44.15% |
| **EfficientNet** | 41.86% |
| **VGG19 + ROI** | 58.87% |
| **Senet_50 + ROI** | 52.87% |
| **Resnet_50 + ROI** | **60.81%** |
| **EfficientNet + ROI** | 46.91% |

## 5. CONCLUSION

In this paper, we propose a simple but very compact method for ME recognition that surpasses other existing research in the field. Feature extraction is considered a critical stage that determines dominant efficiency of any framework. By applying the concept of dynamic image, we take advantage of middle level feature representation and facilitate the problem at hand. Moreover, with the incorporation of deep neural networks, data is exposed to other feature learning layers that makes it more robust and solid. We achieve a high accuracy rate compared to other studies that adopt similar evaluation measuring. However, due to the special traits of ME data, future improvements need to be presented in order to deploy applications in the real world. We suggest that geometric features should also be utilized to represent low-level features in conjunction with high-level representation from deep neural networks. That approach might help classifiers get exposed to better feature representations and hence makes ME systems effective.

## REFERENCES

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., & Gould, S. (2016). Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3034-3042). IEEE.

Davison, A., Merghani, W., Lansley, C., Ng, C. C., & Yap, M. H. (2018, May). 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)* (pp. 642-649). IEEE. doi:10.1109/FG.2018.00101

Davison, A. K., Lansley, C., Costen, N., Tan, K., & Yap, M. H. (2016). Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, *9*(1), 116–129. doi:10.1109/TAFFC.2016.2573832

Ekman, P. (2009). Lie catching and microexpressions. *The Philosophy of Deception, 1*(2), 5.

Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., & Tuytelaars, T. (2015). Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5378-5387). IEEE.

Haggard, E. A., & Isaacs, K. S. (1966). Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy* (pp. 154–165). Springer. doi:10.1007/978-1-4684-6045-2_14

Happy, S. L., & Routray, A. (2017). Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing*, *10*(3), 394–406. doi:10.1109/TAFFC.2017.2723386

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). IEEE.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141). IEEE.

Huang, X., Zhao, G., Hong, X., Zheng, W., & Pietikäinen, M. (2016). Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, *175*, 564–578. doi:10.1016/j.neucom.2015.10.096

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

Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2017). Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, *9*(4), 563–577. doi:10.1109/TAFFC.2017.2667642

Li, X., Pfister, T., Huang, X., Zhao, G., & Pietikäinen, M. (2013, April). A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)* (pp. 1-6). IEEE.

Li, Y., Huang, X., & Zhao, G. (2018, October). Can micro-expression be recognized based on single apex frame? In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3094-3098). IEEE. doi:10.1109/ICIP.2018.8451376

Liong, S. T., See, J., Wong, K., & Phan, R. C. W. (2016, November). Automatic micro-expression recognition from long video using a single spotted apex. In *Asian conference on computer vision* (pp. 345-360). Springer.

Oh, Y. H., See, J., Le Ngo, A. C., Phan, R. C. W., & Baskaran, V. M. (2018). A survey of automatic facial micro-expression analysis: Databases, methods, and challenges. *Frontiers in Psychology*, *9*, 1128. doi:10.3389/fpsyg.2018.01128

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987. doi:10.1109/TPAMI.2002.1017623

Patel, D., Hong, X., & Zhao, G. (2016, December). Selective deep features for micro-expression recognition. In *2016 23rd international conference on pattern recognition (ICPR)* (pp. 2258-2263). IEEE.

Peng, M., Wang, C., Chen, T., Liu, G., & Fu, X. (2017). Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in Psychology*, *8*, 1745. doi:10.3389/fpsyg.2017.01745

Pfister, T., Li, X., Zhao, G., & Pietikäinen, M. (2011, November). *Recognising spontaneous facial micro-expressions. In 2011 international conference on computer vision*. IEEE.

Porter, S., & Ten Brinke, L. (2008). Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, *19*(5), 508–514. doi:10.1111/j.1467-9280.2008.02116.x

Seidenstat, P., & Splane, F. X. (Eds.). (2009). *Protecting airline passengers in the age of terrorism*. ABC-CLIO.

Shahar, H., & Hel-Or, H. (2019). Micro Expression classification using facial color and deep learning methods. *Proceedings of the IEEE International Conference on Computer Vision Workshops*. doi:10.1109/ICCVW.2019.00207

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556

Tan, M., & Le, Q. V. (2019). *Efficientnet: Rethinking model scaling for convolutional neural networks.* arXiv preprint arXiv:1905.11946

Wang, C., Peng, M., Bi, T., & Chen, T. (2020). Micro-attention for micro-expression recognition. *Neurocomputing*, *410*, 354–362. doi:10.1016/j.neucom.2020.06.005

Wang, Y., Ma, H., Xing, X., & Pan, Z. (2020, January). Eulerian Motion Based 3DCNN Architecture for Facial Micro-Expression Recognition. In *International Conference on Multimedia Modeling* (pp. 266-277). Springer. doi:10.1007/978-3-030-37731-1_22

Wu, H. Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics*, *31*(4), 1–8. doi:10.1145/2185520.2185561

Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., Chen, Y. H., & Fu, X. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS One*, *9*(1), e86041. doi:10.1371/journal.pone.0086041

Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 915–928. doi:10.1109/TPAMI.2007.1110

Zhao, Y., & Xu, J. (2019). An improved micro-expression recognition method based on necessary morphological patches. *Symmetry*, *11*(4), 497. doi:10.3390/sym11040497

*Trang Le received her B.A. degree in applied linguistics from Hue University (Vietnam), and is currently pursuing M.Sc degree in software engineering at the University of St. Thomas. Her research interests include facial micro-expression analysis, facial expression recognition, and applied machine learning*

*Khanh Tran graduated with a Bachelor in Mathematics and Computer Science in University of Science, Vietnam (03/2012) and M.Sc. in Electronics Engineering in Chonnam National University, South Korea (08/2015). He is Doctoral Student from Department of Computer Science, University of Oulu, Finland. His research interest are: computer vision, micro-expression analysis, affective computing.*

*Manjeet Rege is an Associate Professor of Graduate Programs in Software and Data Science and Director, Center for Applied Artificial Intelligence at the University of St. Thomas. Dr. Rege is an author, mentor, thought leader, and a frequent public speaker on Big Data, Machine Learning, and Artificial Intelligence technologies. He is also the co-host of the "All Things Data" podcast that brings together leading data scientists, technologists, business model experts and futurists to discuss strategies to utilize, harness and deploy data science, data-driven strategies and enable digital transformation. Apart from being engaged in research, Dr. Rege regularly consults with various organizations to provide expert guidance for building Big Data and AI practice, and applying innovative data science approaches. He has published in various peer-reviewed reputed venues such as IEEE Transactions on Knowledge and Data Engineering, Data Mining & Knowledge Discovery Journal, IEEE International Conference on Data Mining, and the World Wide Web Conference. He is on the editorial review board of Journal of Computer Information Systems and regularly serves on the program committees of various international conferences.*