CLASSIFICATION OF FACIAL MICRO-EXPRESSIONS USING MOTION MAGNIFIED EMOTION AVATAR IMAGES

Ankith Jain Rakesh Kumar, Rajkumar Theagarajan, Omar Peraza and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, Riverside, CA - 92521
{arake001, rthea001, opera002}@ucr.edu, bhanu@cris.ucr.edu

Abstract

Facial micro-expressions are subtle involuntary movements of the facial muscles, characterized by a rapid, short duration and genuine emotions. The detection and classification of these micro-expressions by humans and machines is challenging due to their short duration and subtlety.These micro-expressions have many important applications, especially in therapy, monitoring and depression analysis. It has been shown that during therapy, the facial micro-expressions of patients diagnosed with depression are very difficult to identify and in most cases are very subtle. In this paper, the primary focus is on recognition of facial micro-expressions and to overcome the class imbalance of the datasets. Firstly, a novel approach that uses multiple magnified ratios of Eulerian motion magnification is applied to the videos to extract the suppressed micro-expressions. Secondly, we remove the micro-expression frames with low textural variance and obtain the Emotion Avatar Image (EAI). Finally, Deep Convolutional Neural Network (CNN) is used to extract robust facial features from the motion magnified EAI images. These features are classified into three different classes: positive, negative and surprise. The approach is evaluated on three spontaneous micro-expression datasets SMIC, SAMM, and CASME II, and the results are compared with the current approaches that show the effectiveness and significance of the approach.

1. Introduction

Current Human Machine Interaction (HMI) systems have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. Facial expression, which plays a vital role in social interaction, is one of the most important nonverbal channels through which HMI systems can recognize humans internal emotions. Ekman *et al*. [1]. identified six facial expressions (anger, disgust, fear, happiness, sadness, and surprise) as basic emotional expressions that are universal among human beings. Facial expressions are categorized into two types, namely, facial macro-expressions and facial micro-expressions. Facial micro-expressions are in the form of brief and involuntary facial expressions that appear on a person's face according to the emotions being experienced and last for less than 0.5 seconds [2], [3]. Micro-expressions are very subtle and since they last for less than 0.5 seconds, they are very difficult to detect and usually imperceptible to the human eye [4].

Micro-expressions are rapid, subtle, brief and involuntary facial muscle movements in a real-time scenario. Since these expressions can sustain only less than half a second [5], micro-expression spotting and recognition becomes very difficult for humans and machines. Micro-expression has many potential applications in the field of lie-detection, online-learning, security, health care and game-playing. The micro-expression analysis consists of two main tasks, *spotting* that helps in identifying the micro-expression and *recognition* that aims in identifying the different classes of micro-expressions. The main focus of this paper is to recognize the micro-expression and classify them into different classes of emotions.

Analysis of facial micro-expressions plays a vital role in the field of psychology and is widely used by clinical psychologists and psychiatrists in assessing the mental health of patients. According to a survey done in 2017 [6], 22.1% of Americans aged 18 and older, about 1 in 5 adults suffer from a diagnosable mental disorder. Clinical psychologists treat such disorders by providing therapy that in many cases is equally, if not more, effective than medication. Therapy is a collaborative process between the patient and the clinical psychologist, that helps understand the feelings and emotions of an individual.

As of 2014, there are 106,500 licensed psychologists in the USA and a full time clinical psychologists on average handles 26 patients a week and each therapy session lasts between 45-55 minutes [7]. Based on these statistics, there is a huge demand for clinical psychologists and having an automated tool for detecting and classifying facial micro-expressions will be very beneficial for the psychologists in providing quality healthcare for the patients.

Apart from providing psychological therapy, detecting and analysis of micro-expressions plays a vital role in the treatment of diseases like schizophrenia and autism. Poor social functioning is a disabling feature of schizophrenia. Deficits in facial affect recognition is one feature of its poor functioning, and these have been explored in a number of studies [8], [9], [10]. Many empirical studies [11], [12], [13] have also shown that engaging in one-on-one conversations and activities with autistic patients and observing their micro-expressions helps in understanding their emotions and, thus, provide better care for them. Thus, automatic detection and classification of micro-expressions plays a vital role in assisting clinical psychologists and psychiatrists in providing better healthcare for the patients.

To this end we propose an approach for automatic classification of facial micro-expressions using Convolutional Neural Networks, Eulerian Motion Magnified (EMM) videos, and avatar images. We perform data augmentation by augmenting the training dataset with motion magnified videos with different magnification factors. Experimental results show that by augmenting the training dataset, we are able to improve the classification accuracy and also outperform the state-of-the-art approaches. The rest of this paper is organized as follows. We introduce the related works and our contributions in Section 2. The approach for automatic classification of facial micro-expressions is introduced in Section 3. The experimental results and comparisons with state-of-the-art approaches are presented in Section 4. Finally, Section 5 provides the conclusion.

2. Related Work

Classification of facial expressions is a very important problem and has gained a lot of attention over the past few years [14], [15], but there has been very limited work done for the classification of micro-expressions. Zhao *et al*. [16] used Local Binary Pattern with Three Orthogonal Planes (LBP-TOP) to extract features for classification. LBP-TOP is an extension of Local Binary Pattern (LBP), which helps in distinguishing local texture feature information by translating a vector code into histograms. These histograms are performed on each plane (XY, XT, YT) and finally concatenated into a single histogram feature making it robust to illumination changes.

Davison *et al.* [17], proposed a temporal feature extractor, i.e. 3D Histogram of Oriented Gradient (3DHOG) method, which extracts features in all three directions of motion (XY, XT, YT). Liong *et al.* [18], proposed a new technique to utilize only apex frame to recognize the micro-expression. The feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) is used to enhance the apex frame features.

After the entry of Krizhevsky *et al.* [19] in the Imagenet competition [20], state-of-the-art for feature extraction shifted towards CNNs. Khor *et al.* [21] proposed a method called ELRCN, which uses handcrafted features such as optical flow and optical strain, which is passed to a CNN-LSTM architecture that extracts spatio-temporal features and classified them using Support Vector Machines (SVM).

Peng *et al.* [22] proposed a new approach called Dual Temporal Scale Convolutional Neural Network (DTSCNN), which is a two-stream 3-D CNN model. The two streams of the framework were designed to accommodate different frame-rates of facial micro-expression videos.

Li *et al.* [23] used an approach to detect the apex frames in the frequency domain to recognize the facial micro-expressions and classify them based on the apex frame acquisition. Similarly, Gan *et al.* [24] (OFF-ApexNet) used a divide and conquer approach to identify the apex frame. Based on the acquisition of the apex frame they extracted optical flow features and further classified using CNN. Wang *et al.* [25] used the Eulerian Motion Magnification for recognizing the facial micro-expression.

The main problem in recognizing micro-expressions using CNN models is subtle behavior of micro-expressions which makes it difficult to recognize. To overcome this problem, we use Eulerian Motion Magnification (EMM) to reduce the subtle behavior of facial micro-expressions. The other problems associated in recognizing the facial micro-expressions are lack of large datasets and the unbalanced classes in these datasets that makes it difficult to train CNNs efficiently. We perform data augmentation using different magnitudes of Eulerian Motion Magnification (EMM) [26]. Moreover, by using different magnitudes of motion magnifications (x5, x10, x15) and augment them to the training dataset helps to reduce the problem of unbalance dataset. Furthermore, in order to enhance the appearance of the micro-expression, we use a low intensity expression remover that ignores frames in a video that have very small variation in texture. Next, we compute the Emotional Avatar image (EAI) proposed by Yang *et al*. [27] by performing facial landmark alignment using OpenFace and then averaging all the frames into a single image. The EAI is a spatio-temporal representation of a video sequence that registers the facial features at exact locations and maintains the nonrigid facial muscle movement. The regions of the face that are blurry in the EAI indicate the motion in the video.

2.1. Contributions

- Automatic classification of facial micro-expressions using CNNs and Emotional Avatar Image.
- Performed data augmentation using different magni-

tudes of Eulerian motion magnification to overcome the bias in the dataset.

- Comprehensive evaluation of the proposed approach on three publicly available facial micro-expression datasets.
- Comprehensive cross-dataset evaluation of the proposed approach on three publicly available facial micro-expression datasets.

3. Technical Approach

In this section, we present our proposed approach for facial micro-expression recognition as shown in Fig. 1. In CASME II and SMIC dataset, the cropped faces were available from the sequence of video frames. Since the cropped faces were not available in SAMM dataset, we used the Constrained Local Model (CLM) [28] method and cropped the faces.

3.1. Eulerian Motion Magnification

Eulerian motion magnification [26] exaggerates small motions in videos by incorporating spatial and temporal processing to highlight subtle facial micro-expression in a video as shown in Fig. 2. The videos are first decomposed into spatial frequency bands. The primary goal of processing these bands spatially is to increase the temporal signal-to-noise ratio by spatially applying a low-pass filter to the frames of a video and downsampling the pixels for improving the computational efficiency.

The temporal processing is performed on each spatial frequency band. A bandpass filter is applied to extract the frequency band of interest. Finally, the extracted signal is magnified by a factor of α.

The relationship between temporal processing of bands and Eulerian motion magnification for a given image intensity I(x, t) at position x and time t is expressed as:

$$\hat{I}\{x,t\} = f(x + (1+\alpha)\delta(t)) \tag{1}$$

where δ(t) is the displacement function and α is the magnification factor.

The first-order Taylor series expansion is applied on the image at time t, f(x + δ(t)) about x as:

$$I(x,t) \approx f(x) + \delta(t)\frac{\partial(f(x))}{\partial(x)} \tag{2}$$

Assuming the motion signal δ is within the frequency range of passband of the bandpass filter. Thus, we have

$$B(x,t) = \delta(t)\frac{\partial(f(x))}{\partial(x)} \tag{3}$$

For the general case where $\delta(t)$ may not be entirely within the passband of the temporal filter. Therefore, in this case $\delta_k(t)$, represent the different spectral components of $\delta(t)$. The value of $\delta_k(t)$ will be attenuated by the temporal filtering factor $\gamma_k(t)$. The resulting signal is shown below in Eq.4.

$$B(x,t) = \sum_k \gamma_k \delta_k(t)\frac{\partial(f(x))}{\partial(x)} \tag{4}$$

Solving equation 1, 2, and 3 we get,

$$\hat{I}\{x,t\} = f(x + (1 + \sum_k (1+\alpha_k)\delta_k(t)) \tag{5}$$

where α_k = $\gamma_k\alpha$ is the frequency dependent motion magnification factor and δ_k is the temporal sub-band of the motion signal.

The proposed approach uses motion magnification factor α = 10 to exaggerate the micro-expressions. The value of α = 10 is selected based on [29], as the value of α increases, the distortion and amplification of noise also increases.

3.1.1 Selection of Amplification Factor α

Selection of amplification factor α is very important for the video motion magnification. As the value of α increases, the distortion and amplification of noise also increases which causes artifacts in the video. Fig.3 shows the plot for the Peak-Signal-to-Noise-Ratio Vs the Amplification factor α. From Fig. 3 we can observe that as the amplification factor is increased, the PSNR ratio of the videos rapidly decreases, indicating increasing levels of artifacts being added into the video as the value of α increases. As a result, while testing our approach we use α = 10 for all the testing videos, whereas, the training dataset is augmented with motion magnified videos with α = 5, 10, 15.

3.2. Removing Frames with low Textural Variance

Since micro-expressions are very subtle and last for less than 0.5 seconds, we are only concerned about frames that have high variance in terms of movement of facial muscles [23]. Therefore, it is crucial to get rid off the frames that have very small textural variance.

Similar to [23], in this paper we remove the frames with low textural variance by computing the texture map of individual frames using the Local Binary Patterns (LBP)[30] of individual frames in the video. This texture map is then divided into 6 x 6 blocks. For each block, we compute the frequency to understand the pixel change in the temporal domain and compare the values in the sequence of video frames. Furthermore, we obtain the frequency of each block using 3DFFT with a sliding temporal window size of N which is 61. We mask a sliding window of length N in the current frame, to compute the frequency of frames

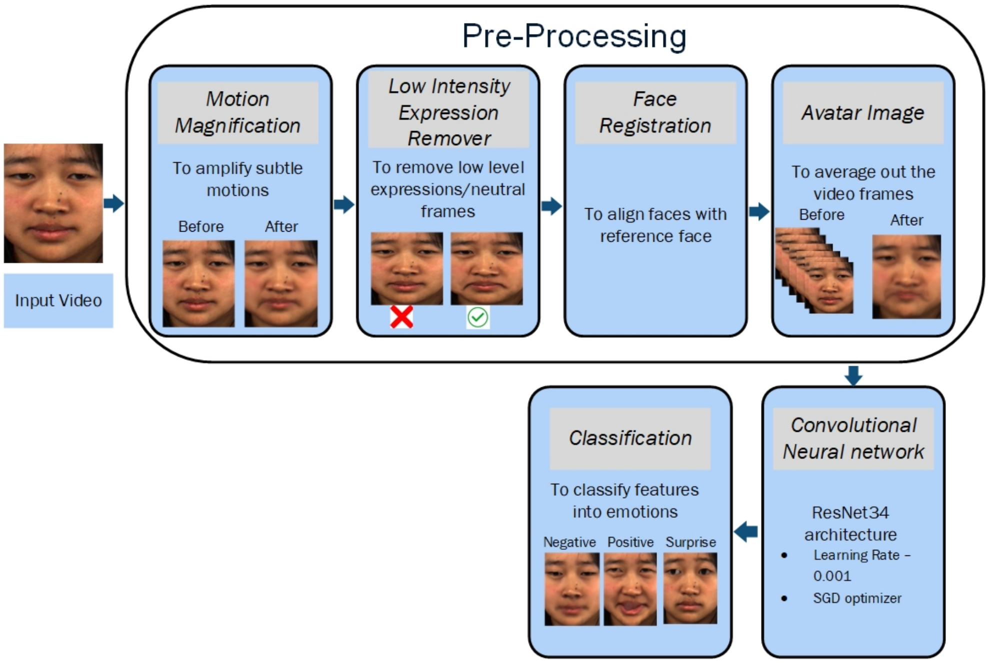
Pre-Processing
Input Video
Motion Magnification
To amplify subtle motions
Before
After
Low Intensity Expression Remover
To remove low level expressions/neutral frames
Face Registration
To align faces with reference face
Avatar Image
To average out the video frames
Before
After
Classification
To classify features into emotions
Negative
Positive
Surprise
Convolutional Neural network
ResNet34 architecture
Learning Rate – 0.001
SGD optimizer

Figure 1. Overall Architecture of our approach.

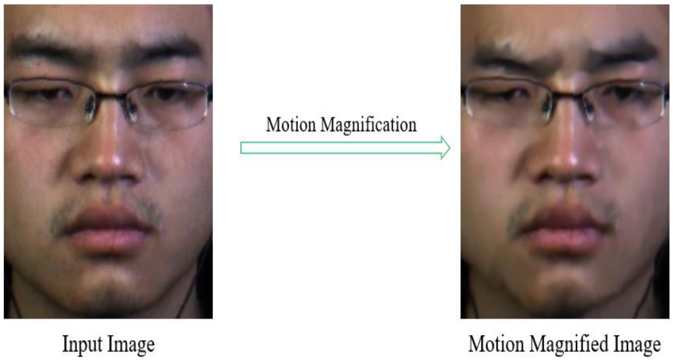

Figure 2. Motion Magnification of an image

present inside the sliding window. We determine the frequency values for the i-th interval on its 36 blocks using 3DFFT. The blocks are represented as (b_{i1}, b_{i2}, ..., b_{i36}). The frequency value for each block in the i-th interval is obtained as:

$$f_{b_{ij}}(x,y,z) = \int_{\frac{-N}{2}}^{\frac{N}{2}} \int_{\frac{-L_b}{2}}^{\frac{L_b}{2}} \int_{\frac{-W_b}{2}}^{\frac{W_b}{2}} F_{b_{ij}}(u,v,q) \times e^{j2\pi(us+vy+qz)} dvdudq \quad (6)$$

where (x,y,z) represents the position in the frequency domain, L_b represents the height of each block; W_b represents as the width of each block, where j = 1, 2, 3, ... 36, in the j-th video interval of block b_{ij}.

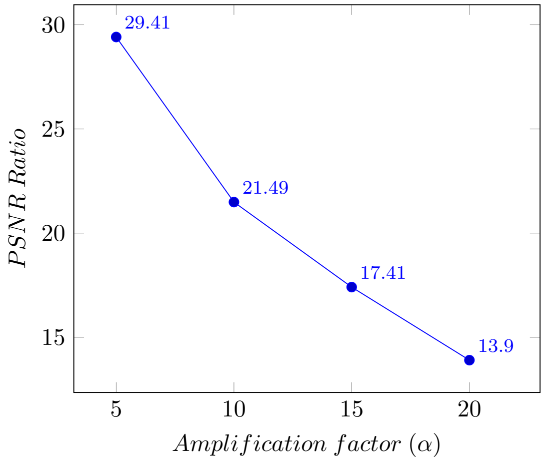

Figure 3. Amplification factor (α) vs PSNR Ratio for video motion magnification

In the micro-expression frames, not all pixel represents the high-frequency values. As a result, we employ a high-frequency band filter (HBF) to remove the low-frequency pixels in the sequence of video frames, such that the removal of these unchanged pixel values helps in getting rid off the insignificant pixel values. The high-frequency filter $H_{b_{ij}}$ is expressed as in Eq.7, where D_o is the threshold value equal to 30.

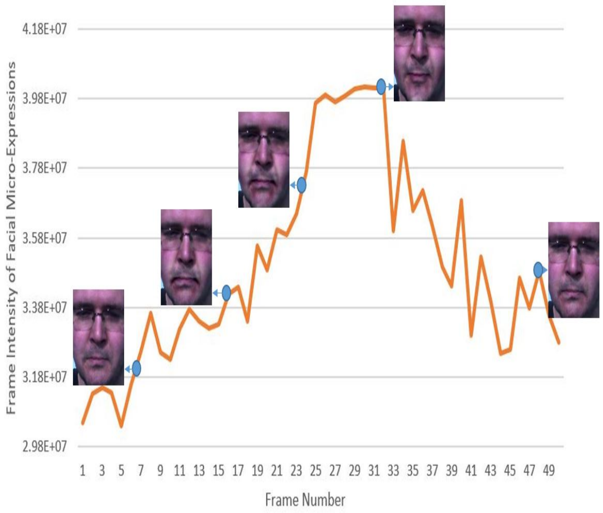

Figure 4. Intensity of Facial Micro-Expression vs Frame Number.

$$H_{b_{ij}}(x,y,z) = \begin{cases} 1, & \sqrt{x^2+y^2+z^2} \geq D_o. \\ 0, & \sqrt{x^2+y^2+z^2} < D_o \end{cases} \tag{7}$$

We filter the video blocks in the frequency domain by using the information from the 3DFFT and high-frequency filter as shown in Eq. 8.

$$G_{b_{ij}}(x,y,z) = f_{b_{ij}}(x,y,z) \times H_{b_{ij}}(x,y,z) \tag{8}$$

Furthermore, we sum the intensity values $G_{b_{ij}}$ for all 36 blocks in the i-th video interval by the Eq. 9.

$$A_i = \sum_{j=1}^{36} \sum_{x=1}^{N} \sum_{y=1}^{L_b} \sum_{z=1}^{W_b} G_{b_{ij}}(x,y,z) \tag{9}$$

where A_i represents the intensity values of each frame of the i-th interval of sequence of video frames. It helps in understanding the changes in the intensity of facial micro-expressions.

Knowing the intensity values for the sequence of video frames from equation 8, we compute the mean and standard deviation for the micro-expression frames and using these values we get the probability values for each frame. We set a threshold value of 0.5 such that if the probability value of a frame is above 0.5, we consider the frame to have high texture variance and the frames which have probability value lesser than 0.5 are removed. The Fig. 4 shows the change in the intensity value of facial micro-expression in a video.

3.3. Facial Landmark Alignment

The input faces can be at a different angle from the camera and can be of different poses. Therefore, in order to compute the Emotion Avatar Image (EAI), it is essential to align the faces with a reference face. In our approach we chose the first frame of the video to be the reference frame. We use the OpenFace software [31] to register the facial images. OpenFace computes the 68 facial landmark points on the facial image and computes an affine transformation to register the face with respect to the reference frame.

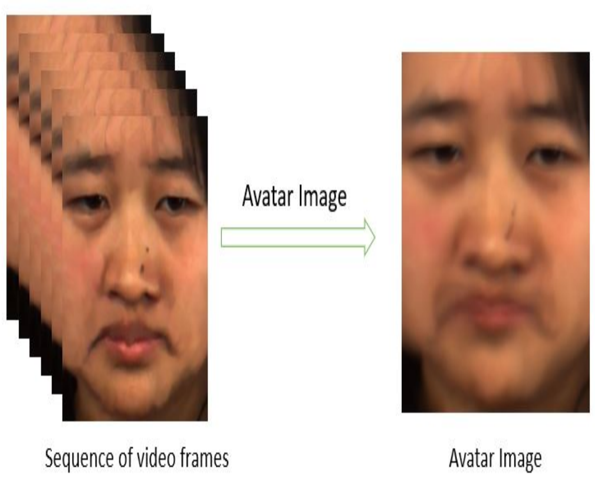

Figure 5. EAI representation for the sequence of frames.

3.4. Emotion Avatar Image

The Emotion Avatar Image (EAI) [27] is a novel method where video sequences are condensed into a single image representation. The technique is simple but effective for facial expression recognition. Here, we use the same approach for facial micro-expression recognition. After registering the individual facial images, the sequence of video frames is averaged out into a single image. The EAI representation registers the facial features at exact locations and maintains the nonrigid facial muscle movement. Therefore, EAI representation helps in highlighting the facial micro-expression in an image. The advantage of using avatar image representation is that it reduces the noise variance by a factor of N, where N is the number of frames in the video. The EAI representation for a sequence of frames is shown in Fig. 5. In Fig. 5 the regions of the face that are blurry indicate facial muscle movement.

3.5. Deep Convolutional Neural Network

We use Convolutional Neural Network, to perform feature extraction. In our approach, we employ the Resnet 34 architecture [32]. The residual network architecture takes an image size of 224x224 and batch normalization is carried out before each convolution layer for faster training convergence. Rectified Linear Unit (ReLu) activation is also used after each convolutional layer.

Network	Learning rate	Momentum	Weight Decay	Optimizer
Resnet 34	10^{-3}	0.9	5 x 10^{-4}	SGD

Table 1. Parameters for the Network

Emotion Class	SMIC	CASME II	SAMM	Combined
Negative	70	88	92	250
Positive	51	32	26	109
Surprise	43	25	15	83
Total	164	145	133	442

Table 2. Summary of the Data Distribution samples.

Table 1 shows the hyper parameters used for training the CNN. We used a mini-batch size of 128 and during every epoch, the training data are randomly flipped and shuffled. The learning rate is decreased after every 5 epochs by a factor of 2. We perform Leave-One-Subject-Out Cross Validation (LOSO-CV) to ensure that each subject is validated and the classification of facial micro-expressions is performed on the emotion avatar image.

4. Experimental Setup and Results

We evaluated our approach on three spontaneous facial micro-expression datasets: SMIC [33], CASME II [34], and SAMM [35] using a Leave-One-Subject-Out Cross Validation approach (LOSO-CV) as shown in Table 2. The framework of our approach is implemented using two NVIDIA GTX 1080Ti GPUs.

4.1. Datasets

4.1.1 SMIC

The SMIC dataset consists of three classes of emotion: Negative (70), Positive (51) and Surprise (43) videos, a total of 164 videos with a frame rate of 100fps. The SMIC dataset consists of 16 subjects.

4.1.2 CASME II

The CASME II dataset consists of seven categories of expressions: other (99), disgust (63), happiness (32), repression (27), surprise (25), sadness (7) and fear (2) in total of 255 videos with a frame rate of 200fps. In our research, we are interested in three classes of expressions: Negative (Disgust and Repression), Positive (Happiness) and Surprise. We chose the three classes based on the rules described [36]. The CASME II dataset consists 24 subjects.

4.1.3 SAMM

The Spontaneous Actions and Micro-Expression (SAMM) dataset consists of eight expressions: anger (57), happiness (26), disgust (9), other (26), fear (8), surprise (15), contempt (12) and sadness (6), in total consists of 159 videos. The frame rate of SAMM dataset is 200fps. In our research, we are interested in three classes of expressions: Negative (Anger, Sadness, Contempt, Fear and Disgust), Positive (Happiness) and Surprise. We chose the three classes based on the rules described [36]. The SAMM dataset consists 28 subjects.

4.1.4 Composite Database Evaluation

The datasets (SMIC, CASME II, and SAMM) are combined into a composite dataset, based on the three emotion classes. In the composite dataset, there are in total of 68 subjects consisting of 442 videos as shown in Table 3. This dataset portrays a real-life scenario consisting of subjects from different ethnicity and gender. The combined datasets consists of bias in the ethnicity based on color of skin and also expression type.

4.2. Evaluation Metrics

The class distributions of these datasets and the composite dataset are imbalanced with respect to the number of classes. Therefore, we cannot use accuracy as the performance metric to gauge our approach. To overcome such imbalance, we use Unweighted F-1 (UF1) score and Unweighted Average Recall (UAR).

- Unweighted F-1 score (UF1): F1 score provides equal emphasis on each class. From the confusion matrix, we compute the True Positives (TP), False Positives (FP) and False Negatives (FN) for each class (C). The balanced F-1 score is obtained by taking the average for each class F1 scores:

$$TP_c = \sum_{i=1}^{k} TP_c^{(i)} \tag{10}$$

$$FP_c = \sum_{i=1}^{k} FP_c^{(i)} \tag{11}$$

$$FN_c = \sum_{i=1}^{k} FN_c^{(i)} \tag{12}$$

$$F1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c} \tag{13}$$

$$UF1 = \frac{F1_c}{C}, \tag{14}$$

Method	Combined Dataset		SMIC		CASME II		SAMM	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
LBP-TOP	0.5785	0.5882	0.5280	0.2000	0.7429	0.7026	0.4102	0.3954
Bi-WOOF	-	-	-	0.6110	-	0.7902	-	0.3970
OFF-ApexNet	0.7033	0.7104	-	0.6817	-	**0.8697**	-	0.5409
Our Approach	**0.7355**	**0.7603**	**0.7621**	**0.7451**	**0.8065**	0.8280	**0.6815**	**0.7056**

Table 3. Comparison of our approach with the state-of-the-art approaches on the three datasets

Class	SMIC		
	Negative	Positive	Surprise
Negative	65	2	3
Positive	11	37	3
Surprise	17	1	25

Table 4. Confusion matrix for the SMIC dataset.

Class	CASME II		
	Negative	Positive	Surprise
Negative	84	2	2
Positive	10	20	2
Surprise	4	0	21

Table 5. Confusion matrix for the CASME II dataset.

Class	SAMM		
	Negative	Positive	Surprise
Negative	85	3	4
Positive	8	17	1
Surprise	6	2	7

Table 6. Confusion matrix for the SAMM dataset.

- Unweighted Average Recall (UAR): It is known as balanced accuracy. Here Acc$_c$ refers to the accuracy per class and n$_c$ refers to the number of items in the class.

$$UAR = \frac{1}{C}\sum Acc_c, \tag{15}$$

where Acc$_c$ = $\frac{TP_c}{n_c}$

4.3. Experimental Results

Table 3 shows the comparison between the current state-of-the-art approaches and our proposed approach using the Leave-One-Subject-Out Cross Validation (LOSO-CV) approach. As a measure of robustness and to handle the class imbalances of the datasets, the performance, and the results are quantified using the balanced metrics: Unweighted F1 score (UF1) and Unweighted Average Recall (UAR). Our proposed approach outperforms the state-of-the-art methods by a huge percentage. The proposed method for data augmentation along with the motion magnified avatar images improves the overall performance of combined dataset and on the individual datasets. The results of SMIC and SAMM datasets of our approach are better than the state-of-the-art methods. This indicates that the addition of data augmented motion magnified avatar images to a class imbalance dataset helps increase the classification accuracy. Table 4, 5, 6 shows the confusion matrix for the datasets.

4.3.1 Cross-Dataset Evaluation

To verify the robustness of our approach and its generalizability to learn the features from different environments and subjects, we use cross-dataset evaluation on the three publicly available Facial micro-expressions dataset.

Table 7 shows the robustness of our approach. We use the same approach as mentioned in the technical approach in section 3. We evaluate our approach on cross-datasets using Leave-One-Subject-Out Cross Validation (LOSO-CV) method. The table compares the performance of our approach on the cross-dataset environment where we train it on one dataset and test it on an other dataset. The results from the cross-dataset evaluation show that the approach is better in generalizing over the large number of subjects.

5. Conclusions

In this paper, we use motion magnified emotion avatar images to highlight the movement of facial micro-expression on the face. The motion magnification helps exaggerate the micro-expression in the avatar image which helps the CNN to classify them into three different classes (Negative, Positive and Surprise). The method is tested on the combined datasets and on individual datasets (SMIC, CASME II, and SAMM) using Leave-One-Subject-Out Cross Validation approach (LOSO-CV). The results from our proposed approach outperforms the current state-of-the-

Training Dataset	Testing Dataset					
	CASME II		SMIC		SAMM	
	UAR	UF1	UAR	UF1	UAR	UF1
CASME II	-	-	0.6684	0.6770	0.6506	0.6595
SMIC	0.6213	0.6442	-	-	0.6076	0.6223
SAMM	0.6044	0.6210	0.5846	0.5924	-	-

Table 7. Cross-Dataset Evaluation for three Facial Micro-Expression Datasets.

art approaches on the UF1 and UAR metrics. We also perform a cross-dataset evaluation of the three publicly available datasets to generalize our approach. The analysis of facial micro-expressions plays a significant role in psychological therapy and the treatment of diseases like schizophrenia and autism. Therefore, these micro-expressions can be used for diverse medical applications.

Acknowledgement

This work was supported in part by Bourns Endowment funds.

References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[2] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *Face and Gesture 2011*, pp. 51–56, IEEE, 2011.

[3] 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, p. e86041, 2014.

[4] W. J. Yan, S. J. Wang, Y. J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, vol. 136, pp. 82–87, 2014.

[5] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, pp. 217–230, Dec 2013.

[6] "Why a clinical psychologist?," `https://www.vapsych.org/why-a-clinical-psychologist-`.

[7] "How many licensed clinical psychologists are there in the usa?," `https://www.apa.org/monitor/2014/06/datapoint`.

[8] G. Sachs, D. Steger-Wuchse, I. Kryspin-Exner, R. C. Gur, and H. Katschnig, "Facial recognition deficits and cognition in schizophrenia," *Schizophrenia research*, vol. 68, no. 1, pp. 27–35, 2004.

[9] C. G. Kohler, T. H. Turner, W. B. Bilker, C. M. Brensinger, S. J. Siegel, S. J. Kanes, R. E. Gur, and R. C. Gur, "Facial emotion recognition in schizophrenia: intensity effects and error pattern," *American Journal of Psychiatry*, vol. 160, no. 10, pp. 1768–1774, 2003.

[10] T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," *British journal of clinical psychology*, vol. 45, no. 4, pp. 579–583, 2006.

[11] E. B. Torres, M. Brincker, R. W. Isenhower III, P. Yanovich, K. A. Stigler, J. I. Nurnberger Jr, D. N. Metaxas, and J. V. José, "Autism: the micro-movement perspective," *Frontiers in integrative neuroscience*, vol. 7, p. 32, 2013.

[12] C. Tardif, M.-H. Plumet, J. Beaudichon, D. Waller, M. Bouvard, and M. Leboyer, "Micro-analysis of social interactions between autistic children and normal adults in semi-structured play situations," *International Journal of Behavioral Development*, vol. 18, no. 4, pp. 727–747, 1995.

[13] T. F. Clark, P. Winkielman, and D. N. McIntosh, "Autism and the extraction of emotion from briefly presented facial expressions: stumbling at the first step of empathy.," *Emotion*, vol. 8, no. 6, p. 803, 2008.

[14] Z. Wang, Q. Ruan, and G. An, "Facial expression recognition using sparse local fisher discriminant analysis," *Neurocomputing*, vol. 174, pp. 756 – 766, 2016.

[15] A. T. Lopes, E. D. Aguiar, A. F. D. Souza, and T. O. Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610 – 628, 2017.

[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, pp. 915–928, June 2007.

[17] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *CoRR*, vol. abs/1708.07549, 2017.

[18] S. Liong, J. See, K. Wong, and R. C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82 – 92, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, (USA), pp. 1097–1105, Curran Associates Inc., 2012.

[20] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009.

[21] H. Khor, J. See, R. C. W. Phan, and W. Lin, “Enriched long-term recurrent convolutional network for facial micro-expression recognition,” in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 667–674, May 2018.

[22] 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.

[23] 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)*, pp. 3094–3098, Oct 2018.

[24] Y. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan, “OFF-ApexNet on micro-expression recognition system,” *Signal Processing: Image Communication*, vol. 74, pp. 129 – 139, 2019.

[25] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, “Effective recognition of facial micro-expressions with video motion magnification,” *Multimedia Tools Appl.*, vol. 76, pp. 21665–21690, Oct. 2017.

[26] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, “Eulerian video magnification for revealing subtle changes in the world,” *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, vol. 31, no. 4, 2012.

[27] S. Yang and B. Bhanu, “Understanding discrete facial expressions in video using an emotion avatar image,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, pp. 980–992, Aug 2012.

[28] T. Baltruaitis, P. Robinson, and L. Morency, “3d constrained local model for rigid and non-rigid facial tracking,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2610–2617, June 2012.

[29] A. C. L. Ngo, A. Johnston, R. C. . Phan, and J. See, “Micro-expression motion magnification: Global lagrangian vs. local eulerian approaches,” in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 650–656, May 2018.

[30] D. C. Li and L. Wang, “Texture unit, texture spectrum, and texture analysis,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, pp. 509–512, July 1990.

[31] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “Openface: A general-purpose face recognition library with mobile applications,” tech. rep., CMU, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[33] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikinen, “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, April 2013.

[34] 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, pp. 1–8, 01 2014.

[35] 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, pp. 116–129, Jan 2018.

[36] “Second micro-expression grand challenge (megc).” https://facial-micro-expressiongc.github.io/MEGC2019.