Micro-Expression Recognition using 3D - CNN

Vishal Dubey, Bhavya Takkar, P. Singh Lamba *

Bharati Vidyapeeth's College of Engineering, INDIA

Emails: dubeyvishal1998@gmail; bhavyat.work@gmail.com; singhs.puneet@gmail.com
* Correspondence: singhs.puneet@gmail.com

Abstract

Micro-expression comes under nonverbal communication, and for a matter of fact, it appears for minute fractions of a second. One cannot control micro-expression as it tells about our actual state emotionally, even if we try to hide or conceal our genuine emotions. As we know, micro-expressions are very rapid, making it challenging for any human being to detect them with bare eyes. This subtle expression is spontaneous and involuntary and gives an emotional response. It happens when a person wants to conceal a specific emotion, but the brain reacts appropriately to what that person is feeling then. Due to this, the person displays their true feelings very briefly and later tries to make a false emotional response. Human emotions last about 0.5 - 4.0 seconds, whereas micro-expression can last less than 1/2 of a second. On comparing micro-expression with regular facial expressions, it is found that for micro-expression, it is complicated to hide responses to a particular situation. Micro-expressions cannot be controlled because of the short time interval, but with a high-speed camera, we can capture one's expressions and replay them slowly. Over the last ten years, researchers from all over the globe have been researching automatic micro-expression recognition in computer science, security, psychology, and many more. This paper aims to provide insight regarding micro-expression analysis using 3D CNN. A lot of datasets of micro-expression have been released in the last decade. We have performed this experiment on the SMIC micro-expression dataset and compared the results after applying two different activation functions.

Keywords: 3D CNN; micro-expression; Micro-Expression Recognition

1. Introduction

Facial expression is a common way to convey our emotions, it has a very long history in the field of research, and from the 1970s, it accelerated quickly. Modern theory by Ekman et al. [1], [2], [3] on primary emotion has given a whole new dimension to the psychology of emotion [4]. The seven universal facial expressions are contempt, happiness, sorrow, anger, fear, surprise, and disgust. When somebody is emotional, there is an impulse reaction according to the situation, which tells the state of emotion of that person.

Micro-Expression duration is very less. That is why micro-expression analysis has become such an essential aspect of research. It occurs when a person tries to hide their true feelings; the origin of subtle expression happens in this scenario [2], [5]. Facial expression can give away the state of emotion of any person, which might not be appropriate in some cases due to the cultural code of conduct [6]. That person may try to suppress the facial expression. Once subdued, the person can mask the original facial expression and cause subtle feelings, which is much more suitable for the situation. Identifying a micro-expression is a very challenging and challenging task for any human being.

Micro-expressions are involuntary and contain significant amounts of true feelings that are useful in the real world and help in interrogations and security-related aspects. [7], [8], [9]. This information is not easy to extract because there are fast movements in micro-expressions, where this is essential that features should be more descriptive. One of the main features of micro-expressions is their short duration, which by normal standards does not exceed 500 ms

Doi : https://doi.org/10.54216/FPA.010101

Received: January 25, 2020 Revised: March 28, 2020 Accepted: April 15, 2020

[10]. In various other studies, the speed of micro-expression has been discussed, and according to that, some say it occurs for less than 250 ms [11], some say less than 330 ms [3], and at least it can be less than half a second [8]. Micro-expression is very rapid, due to which we cannot identify the genuine emotion that person is feeling. In recent years many researchers have researched various methods helpful for micro-expression recognition, especially deep learning methods[12].

Micro-expression recognition is very helpful in critical situations like an interrogation of any suspect or a matter of public security. It has numerous applications, and it is nowadays considered the most desired area in the field of research[13]. The most popular deep learning approach widely used for image classification is CNN (Convolutional neural network) [13,14,15]. Various strategies are there which help us to recognize micro-expressions; these approaches are CNN based and are helpful for still images. Spontaneous micro-expression video is very much difficult to tackle because, in its processing, we require multiple frames for both temporal and spatial information. Two different 3D CNN models were proposed earlier to extract the information over the video, and on two datasets, this experiment was conducted [16]. In this paper, we are using deep learning approaches for micro-expression recognition on the SMIC [21] dataset, which is a type of spontaneous dataset.

2. Methodology

2.1. Facial Micro-Expression Dataset

They are two types of facial micro-expression datasets available, which are furthermore classified into various other datasets [12]:

2.1.1.Non-spontaneous Dataset

2.1.1.1 Polikovsky Dataset [17]
2.1.1.2 USF-HD [18]
2.1.1.3 York DDT [19]

2.1.2 Spontaneous Dataset

2.1.2.1 Chinese Academy of Sciences Micro-Expression [20]
2.1.2.2 Spontaneous Micro-expression Database [21]
2.1.2.3 Chinese Academy of Sciences Micro-Expression II [22]
2.1.2.4 Spontaneous Action and Micro-Movements [23]
2.1.2.5 A database of Spontaneous micro-expression and macro-expression [24]

This paper recognizes micro-expression using 3D CNN on SMIC [21] dataset.

SPONTANEOUS MICRO-EXPRESSION DATABASE (SMIC)

The study of micro-expression is a difficult task, and the creation of a spontaneous micro-expression dataset is also very complicated. Still, Le at el. [21] created the SMIC dataset. In this dataset, 20 participants were included, and the location of setup was an indoor environment, which was designed in such a way that it resembled an interrogation room. High speed (HS) camera was used for recording the first 10 participants, and for the latter 10 participants, they integrated two more cameras, NIR (Near – Infrared) and VIS (Normal Visual). HS camera of 100 fps and NIR, VIS were of 25 fps. The resolution of all the three cameras was the same 640 × 480. The researchers left the participants alone in that room, and they recorded their reactions of the participants. Later, 164 clips of 16 participants were finalized for the SMIC database, and these clips were recorded in the HS dataset. The three classes in this dataset are positive, negative, and surprise. We have made some alterations in the dataset and refined it a bit in the out model; there is a total of 158 inputs and three classes in which they are classified. Out of those 158, 66 inputs are labeled as negative, 50 are labeled as positive, and 42 are labeled as a surprise.

Doi : https://doi.org/10.54216/FPA.010101

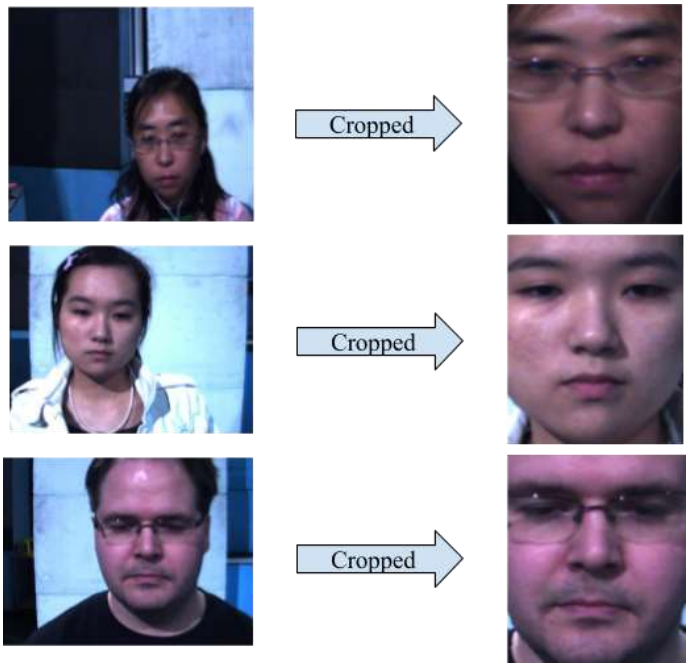

Figure 1: SMIC dataset cropped images (Sample from Negative, Positive, and surprise classes, respectively)

As shown in fig.1, the cropped face portion is passed as input to the model for preprocessing. This helps achieve the result more easily as only the points most important for our problem statement are taken. Labels are given to the classes in this micro-expression dataset, as shown in Table I.

Table 1: Labels given to different classes

CLASSES	LABEL
Negative	0
Positive	1
Surprise	2

2.2 Performance parameters

The confusion matrix or error matrix is most commonly used in machine learning, especially for classification problem categories. An error matrix or confusion matrix describes the performance of the classification model on a set of samples that are kept for testing purposes for which the actual values are known to us. It helps us visualize the working or performance of a model or an algorithm. Generally, in a classification problem, there can be confusion between two or more classes when a sample is being classified into one of its classes, or in simpler words, we can say that it can be mislabelled as the other class's object. There are various performance measures that we can compute with the help of a confusion matrix.

The key to the confusion matrix is the classes that are present in it. When a confusion matrix is made for the data set, it is placed in the correct predicted or incorrectly predicted class with its count values in it. When data is classified, it can be classified into other classes in which it doesn't belong; this is called mislabelling or confusion. So, the confusion matrix shows us the errors our model or algorithm is making and what type of errors it is making.

Doi : https://doi.org/10.54216/FPA.010101

There are two classes:

- Class 1: Positive
- Class 2: Negative

Definition of the terms used in the confusion matrix (for two classes, it can be extended for multiclass as well):

- Positive (P): The sample is positive (e.g., A cat)
- Negative (N): The sample is negative (e.g., a dog)
- True Positive (TP): The sample is a cat and is predicted to be a cat.
- False Positive (FP): The sample is a dog but is predicted to be a cat.
- True Negative (TP): The sample is a dog and is predicted to be a dog.

Performance measures used to classify micro-expressions

1. Accuracy of classification rate

   It is given by the relation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. Recall

   It is defined as the ratio of the total number of correctly classified positives to the total number of positive samples.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

3. Precision

   It is defined as the ratio of the total number of correctly classified positive samples to the total number of positive classified samples.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4. F-score

   It is basically the harmonic mean of recall and precision. Its value will always be nearer to the smaller recall or precision.

$$F - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

2.3 Activation Function

Mathematical equations can determine the output of the Neural network, and these equations are also known as the activation function. With each and every neuron, this function is attached and determines with the input provided by the neuron whether it should be activated or not. The Neuron output can also be normalized with the help of the activation function. The range can be between -1 to 1 or 1 and 0. The choice of activation function in deep neural networks has a significant amount of impact on the training and performance of the working model. The most

Doi : https://doi.org/10.54216/FPA.010101

widely used activation function, inarguably is ReLU activation function, which is f(x) = max (0, x) as shown in figure 2. It looks very simple but does wonder when used in deep neural networks for normalizing the output and helps in decreasing the computing pressure of the model. But there are certain problems in the ReLU activation function, which certainly do not overshadow its advantage. Google brain team proposed a new activation function, named Swish, which is f(x) = x*sigmoid(x) as shown in figure 3. In many of the experiments, it has been seen that Swish works slightly better than the ReLU activation function. It outperforms the ReLU activation function by a considerable amount when incorporated in deep neural networks, but in external neural networks, the results are comparable. In this paper, results are compared using ReLU and Swish activation functions in the model.

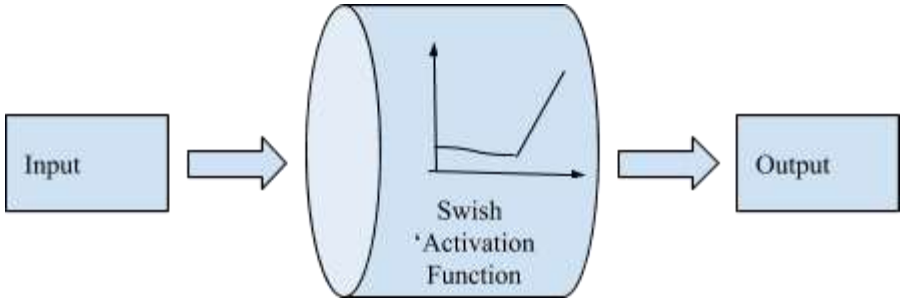

Figure 2: SWISH Activation Function

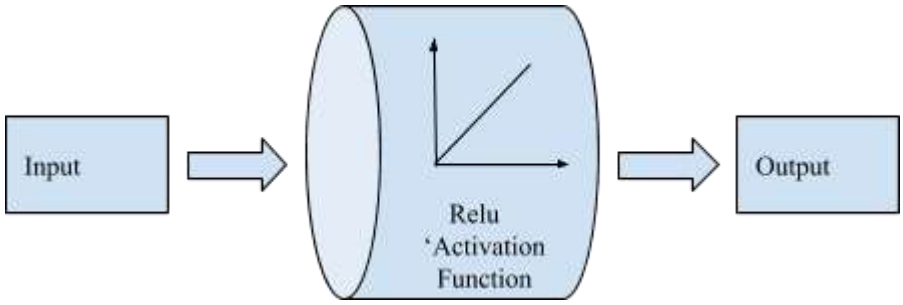

Figure 3: RELU Activation Function

3. Experimental Setup

Micro-expression recognition is a very tedious task. First of all, it is not very easy to gather the dataset for recognition, and second, there are not many approaches to do that. The problem arises because of its certain properties, which make it different from basic human emotions. Micro-expression is very hard to catch with the naked eyes because of its occurrence in a very little time span. It generally occurs for half a second or sometimes even less than that, due to which it is very hard to gather the data for the micro-expression. So, for our model, we took the help of the SMIC HS dataset for the purpose of micro-expression recognition. The SMIC dataset has three output classes named Positive, Negative, and Surprise.

There are various approaches to micro-expression recognition, which include some of the standard algorithms such as the Three-dimensional Histogram of Gradients (3D-HOG) and Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [25]. A lot of work has also been done using the Deep Learning approach. As we can see, the term "Three" is used very frequently whenever we are discussing the approaches of Micro-expression recognition because we require a set of images for input in our model as we cannot determine the micro-expression of people by a single image because it occurs for a very minute amount of time. It can be determined when we take a group of images together, and more images lead to better visualization. As we have already talked about, there are various approaches to recognizing facial micro-expression, which include deep learning.

Doi : https://doi.org/10.54216/FPA.010101

The Convolutional Neural Network (CNN) is the most frequent term which we hear whenever Image Processing is being discussed. CNN has the ability to fetch those very small details, which makes it many steps ahead of other approaches. There's the convolutional layer that convolutes the input set of images with a filter, which gives rise to a newly generated matric with the most prolific features, which could be more helpful in deciding the class of the images. There is also a pooling layer which takes the most promising feature of the classification process. To remove the extra features which most probably will not affect the classification can be dropped as well with the help of the dropout layer. It also decreases the computation needed for this tedious task as there can be millions and millions of input variables needed to be computed and processed. This affects the efficiency of the model as well, as sometimes it takes even days to compute the result of the model in Neural networks. There are various activation functions that will help reach the most optimized weights and bias for the input variables to classify the inputs into the right classes. So, CNN has proved itself very helpful in image processing; that's why we decided to go with the CNN approach and perform micro-expression recognition. We have used seven different layers in our 3D-CNN model shown below in fig 4. All the layer has different functions, which we have explained above.

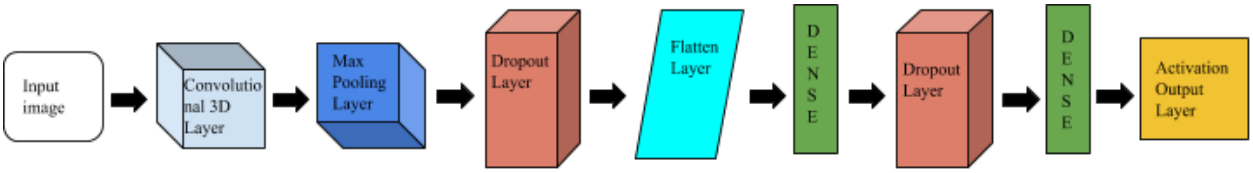

Figure 4: Layers of 3D-CNN models

We used three-dimensional CNN for feature extraction and classification as well. The slight difference between 2-D CNN and 3-D CNN is the kernel. The kernel in 2-D CNN the 2-dimensional, and it only moves in two directions. And on the other hand, the kernel in 3-D CNN is 3-dimensional, and it shifts in the three directions. As we have discussed earlier that it is more efficient and easy to determine the micro-expression when there is a set of continuous images, that's why we used 3D-CNN so that we can take 18 frames of the images together and use that as an input for our model and apply 3-D CNN on that.

Insights of our model

- The images are converted from RGB to Grayscale and then processed.
- The images are of resolution 64x64x18.
- The labels are hot-encoded labels to make it easy for the computation which occurs inside the neural network.
- We used two activation functions: a) RELU b) SWISH
- We also divided the training and testing data into 80 to 20, 70 to 30, 60 to 40, and 50 to 50 ratios to compare the results.

4. Results

In this section, we are presenting our experimental results. We have trained and validated our model on a different set of samples. There are 4 sets of samples we have taken for training and validating purposes:

- Training sample: 80% and Validating sample: 20%
- Training sample: 70% and Validating sample: 30%
- Training sample: 60% and Validating sample: 40%
- Training sample: 50% and Validating sample: 50%

Our model's basic performance measure is accuracy, the loss is categorical cross-entropy, and the Optimizer is SGD. Two activation functions are used in this 3D-CNN model, and we have also compared their results. We have shown the comparison of accuracies for a different set of samples, the Confusion matrix for the best-achieved accuracy, and the the ROC curve for the same. Table II lists the accuracy achieved by the 3D CNN model using two activation functions, ReLU and Swish. It also lists the results when the model is trained on a different set of samples and validated on a different set of samples. From the Table, it can be seen that better results are obtained when more data is used for training the model. Pre-processing is another thing, but the number of samples also affects the model by a very certain amount. We also have a confusion matrix to better understand the result. The first row of the confusion

Doi : https://doi.org/10.54216/FPA.010101

matrix is for the negative class, the second row is for the positive class, and the last row is for the surprise class. Figure 5 and 6 shows the confusion matrix for ReLU and SWISH model.

Table 2: Result Of Diffirent Activation Function

	RELU				**SWISH**			
TRAINING/TESTING RATIO	80/20	70/30	60/40	50/50	80/20	70/30	60/40	50/50
TRAINING ACCURACY	67	69.9	64.8	67	68	63.6	61.7	64.5
VALIDATION ACCURACY	62.5	54.1	42.1	48.1	62	56.2	47.7	53.1

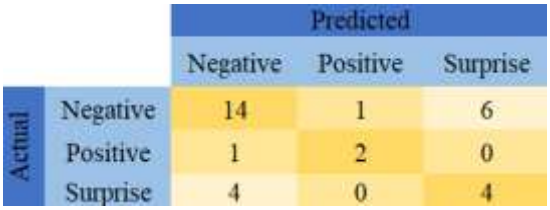

Figure 5: Confusion Matrix for ReLU

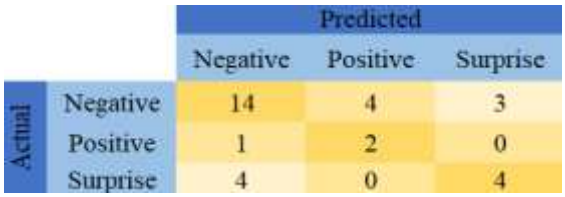

Figure 6: Confusion Matrix for SWISH

Table 3: Performance Measures Of the ReLU Model

CLASS	**N (TRUTH)**	**N(CLASSIFIED)**	**ACCURACY**	**RECALL**	**PRECISION**	**F1 SCORE**
0 (Negative)	21	19	62.5%	0.67	0.74	0.70
1 (Positive)	3	3	93.75%	0.67	0.67	0.67
2(Surprise)	8	10	68.75%	0.50	0.40	0.44

Table 4: Performance Measures Of the Swish Model

CLASS	**N (TRUTH)**	**N(CLASSIFIED)**	**ACCURACY**	**RECALL**	**PRECISION**	**F1 SCORE**
0 (Negative)	21	19	62.5%	0.67	0.74	0.70
1 (Positive)	3	6	84.38%	0.67	0.33	0.44
2(Surprise)	8	7	78.13%	0.50	0.57	0.53

The performance measures we used to analyze our micro-expression recognition result using the 3d CNN model in which we incorporated Swish and ReLU activation functions are shown in Table III and Table IV. It shows the

Doi : https://doi.org/10.54216/FPA.010101

ability of the model to classify rate (Accuracy), recognize the class (Recall), the ability of the model to classify the objects into classes in which they really belong (Precision), and the F1 score, which is defined as the harmonic mean of recall and precision. We can see from the Table that ReLU activation functions overtake the Swish activation function used in our models by a considerate amount.

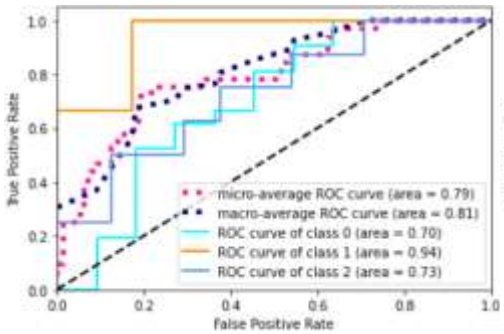

Figure 7: ROC curve for Swish

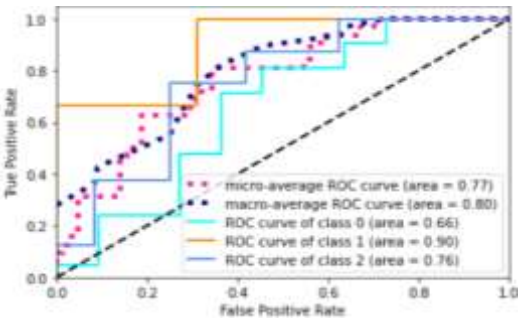

Figure 8: ROC curve for ReLU

ROC curve is a performance measurement for classification problems. ROC (Receiver Operating Characteristics) is a probability curve, and AUC (Area Under the Curve) represents the ability of the model to separate between classes. It basically represents the extent or degree, or measure of separability. The higher the area under the curve, the better the model is at classifying the samples into their respective classes. As we can see from fig. 7 & fig. 8, the area under the curve for Class 0 (Negative) is 0.70 and 0.66 for ReLU and Swish, respectively, which means that the ReLU model is better at separating the objects of the negative class as compared to the Swish model. Similarly, the ReLU model is more capable of distinguishing the objects of the positive class (Class 1) as compared to the Swish model, but the Swish model slightly overtakes the ReLU model in distinguishing samples of class 3 (Surprise).

5. Conclusion

We have proposed 3D-CNN in this paper as a potential approach for micro-expression recognition. We have performed experiments using the SMIC dataset, a spontaneous facial micro-expression dataset. As we know that there are three classes in the dataset, we have labeled these classes, converted the images into grayscale, and incorporated seven layers into our model. In the last layer, we have used two activation functions, and we have also compared the results for the same. We have also made the confusion matrix and Roc curve for the best result that we got from those activation functions. On the basis of results and discussions, we conclude that deep learning approaches are better and give high accuracy.

References

[1] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, pp. 169–200, 1992.
[2] Paul Ekman, Emotions Revealed: Understanding Faces and Feelings. *Phoenix*, 2004.
[3] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, ser. Series in Affective Science. Oxford University Press, 2005.
[4] J. A. Russell and J. M. Ferna'ndez-Dols, *The psychology of facial expression.* Cambridge university press, 1997.
[5] P. Ekman, "*Lie catching and microexpressions*," in The Philosophy of Deception, C. W. Martin, Ed. Oxford University Press, 2009, pp. 118–133.
[6] D. Matsumoto, S. H. Yoo, and S. Nakagawa, "Culture, emotion regulation, and adjustment." *Journal of personality and social psychology*, vol. 94, no. 6, p. 925, 2008.
[7] M. O'Sullivan, M. G. Frank, C. M. Hurley, and J. Tiwana, "Police lie detection accuracy: The effect of lie scenario." *Law and Human Behavior*, vol. 33, no. 6, p. 530, 2009.
[8] M. G. Frank, C. J. Maccario, and V. l. Govindaraju, "*Behavior and security*," in Protecting airline passengers in the age of terrorism. Greenwood Pub. Group, 2009.

Doi : https://doi.org/10.54216/FPA.010101

[9] M. Frank, M. Herbasz, K. Sinuk, A. M. Keller, A. Kurylo, and C. Nolan, "I see how you feel: Training lay people and professionals to recognize fleeting emotions," *in International Communication Association*, 2009.
[10] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.
[11] P. Ekman, Telling Lies: *Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton, 2001.
[12] W. Merghani, A. K. Davison, and M. H. Yap, "A review on facial microexpressions analysis: Datasets, features and metrics," *arXiv preprint arXiv:1805.02397*, 2018.
[13] J. Li , Y.Wang,J.See,W.Liu , "Micro-expression recognition based on 3D flow convolutional neural network" in *Springer-Verlag London*,2018.
[14] He K, Zhang X, Ren S, Sun J," Deep residual learning for image recognition." *arXiv preprint arXiv :1512.03385* , 2015.
[15] Zeiler MD, Fergus , "*Visualizing and understanding convolutional networks."* In Comper vision–ECCV 2014, Springer, pp 818–833
[16] S P.T. Reddy, S T.Karri, S R. Dubey, S. Mukherjee," Spontaneous Facial Micro-Expression Recognition using 3D spatiotemporal Convolutional neural networks",2019
[17] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in Crime Detection and Prevention (ICDP 2009), 3rd International Conference on. IET, 2009, pp. 1–6.
[18] M.Shreve,S.Godavarthy, D.Goldgof,andS.Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), 2011, pp. 51–56.
[19] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 59–69, 2009.
[20] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces," in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013, pp. 1–7.
[21] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in Automatic Face and Gesture Recognition(FG),2013 10th IEEE International Conference and Workshops on. IEEE, 2013, pp. 1–6.
[22] 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.
[23] 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, Jan 2018.
[24] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas (me)ˆ2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, 2017.
[25] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, pp. 915-928, June 2007.

Doi : https://doi.org/10.54216/FPA.010101