Recognizing facial expressions from videos using Convolutional Neural Networks (CNN) and Feature Aggregation

Ratnalata Gupta[1], Prof. (Dr.) L. K. Vishwamitra[2]
Oriental University, Indore, India, Ratnalata@gmail.com, lkvishwamitra@gmail.com

Abstract:

The facial expression recognition (FER) system has always been in trend and majorly when it comes to the FER from the video clips, then it becomes a crucial task to do. The differences seen in the visual descriptor and the emotions seen on the face need to be filled in the FER from videos. Here in our proposed work aggregation is done between the spatial and temporal convolutional features available in the whole video to recognize facial expressions in any video. We are using here 15-15 both spatial and temporal streams. Every stream, i.e. spatial, corresponds with the temporal flow, which creates a layer of aggregation for end-to-end FER system training from the video. This training gives a better representation of the video and avoids overfitting from the limitation of available datasets. We have found that the proposed approach is best for aggregating the video dataset's spatial-temporal features compared to other available methods. The dataset we have used are RML, MMI, BAUM-1s, FER-2013 and the results obtained after applying the proposed approach on this datasets are satisfactory.

Keywords: Video - Face Expression Recognition, Feature Aggregation, CNN, Spatial-Temporal Features, Max-Pooling Layer.

DOI Number: 10.14704/nq.2022.20.13.NQ88332 Neuro Quantology 2022; 20(13):2653-2661

I. INTRODUCTION

Facial Expression Recognition system has come in recent times as the interaction between the human and computer machine has increased tremendously and has a wide range of practical life applications [1]. Video-based FER has been started working, and the classification of the expressions on humans' faces is done through this system. Video-based a large number of researchers has studied FER systems as this system faces a significant problem of fitting the gap seen in the visual features in the video and the emotions seen on face in videos [2]. Wide range of video-based FER application can be seen in industries like robotics, medical sector, driver safety etc. [3]. The primary facial expressions that can be seen classified are happy, fear, sad, surprised, disgust, angry [4].

The facial expression recognition system can be seen in two types, and it can be classified based on the representation of the features [5]. The first classification is based on the static images, and the facial expressions are recognized from this spatial information present in the static images. It seems to be an easy task as compared to the second classification of FER system. The second type is the video-based FER in which the facial expression is recognized from the video clips or real-time videos. This type of FER is considered as the typical work to carry out as in this the work is done in between the two adjacent video frames and the temporal information in between these two frames is carried out So this work deals with the spatial and temporal information present in the video sequence and processing it in place of the standard static images. The video-based FER works in three stages. The first one is video pre-processing, forwarding with the extraction of visual features.

At last, the recognition of expression is done [6]. Cropping of images of the face is done in the pre-processing video stage, and the video frame facial images are used for the extraction of visual

features. In the end, the classification algorithm is applied for the recognition of expression.

Some of the deep learning methods like CNN have been used to learn the video's facial expression recognition [7]. CNN provides higher accuracy for the classification task for the massive video datasets. The datasets are having noise and are available in less quantity. The information like spatial and temporal is useful for the modelling of the videos having different emotions. Recently the work in video-based FER has been carried out based on the spatial and temporal features and 3D convolutions. The conditions seen in 3D convolution are that there is difficulty in scaling it, whereas the spatial and temporal can perform only short videos. So the exploitation of the comprehensive feature has been an essential key for the FER from the video.

Here in our paper, we have presented an approach using CNN framework that is trained end to end for the aggregation and extraction of the descriptors present in the static images and in between the ephemeral streams of video. Here in the core of the descriptors, we have implemented a spatial-temporal aggregation layer named EmotionalVLAN working for the classification in the static images we have and recognizing the videos' actions. The various segments of the CNN model help in the aggregation of features and the FER task are carried out through the penultimate layer on the CNN model. We have also implemented various spatial and temporal information fusion methods and the aggregation of this information into the single video level descriptors.

II. LITERATURE SURVEY

The FERS area of work has expanded to include different fields of expertise in recent years through an automated computerized system. Several researchers have worked to solve the issue of recognizing facial expressions. The goal of this research is to develop an automated system for recognizing facial expressions in real time. Villanueva, M. G., & Zavala, S. R. (2020) [8] In real-time, this article constructs a deep neural network to evaluate only two emotions -- happy and sad -- dataset created by the author in which altering pose, age, and expression. In noisy data, the proposed architecture has achieved 90% accuracy.

Agarwal, S. et al. [9] have designed a system that operates at 10 frames per second on computers and at 2 frames per second on android mobile devices. Face expressions are analyzed based on entropy and correlation. In order to extract facial expressions, they focused only on salient features. According to the authors, their accuracy increased by 13% to 20% compared to other methods.

J. Chen et al. [10] have designed another effective strategy for FER. It is used for facial images as well as speeches from the video. The HOG-TOP feature descriptor has been used to extract facial expressions. CK+ and AFEW 4.0 were used in the experimental analysis.

Y. Ding et al. [11] The purpose of this paper is to present a method for recognizing facial expressions from real-time video. DLBP was presented by the authors as a method for extracting peak expression frames from videos. Here, facial expressions were also obtained using the Laplace logarithm. For accurate detection, JAFFE and CK datasets were used with Taylor feature patterns. It is compared with some state-of-the-art approaches currently in use.

H. Kabir et al.[12] The authors have presented the Oriented Motion Flow method for FER using video patterns. Since POMF has already studied directional motions, the micro-patterns showed in the video help to further improve it. The POMF histogram has been used to implement a hidden Markov Model. A video-based RGB and depth camera-based evaluation is conducted to evaluate the performance of the proposed methodology.

K. S. Yadav[13] Facial expression is one of the most important tools for communication. The authors have been working on the fast-tracking algorithm of Viola-John's. A KNN classifier has been used to classify the emotions recognized. It was decided to evaluate the proposed approach using four databases.

S. Liu et al.[14] the authors should use videos from YouTube or mobile devices to identify the facial expressions. The landmark detection of the 2D images was performed using supervised global descent techniques. It has been examined whether dense reconstruction is possible for

YouTube videos and other sources. Different lighting conditions did not affect the system's ability to track facial landmarks.

P. I. Rani and K. Muneeswaran [15] This paper describe facial expression recognition using the eyes and mouth regions. Gabor wavelets are used in this case to extract the mouth and eye regions from the provided video. An ensemble classifier is used to detect the eye and mouth regions. The Multiclass Adaboost technique has been used to identify facial emotions. CK and RML databases are used in the experiment.

S. Zhang[16] This paper presents a hybrid deep learning model based on two CNNs. This CNN is used for both spatial and temporal features. The outputs of these two CNNs are integrated using a deep belief network. An SVM classifier is used to categorize facial expressions.

III. PROPOSED METHODOLOGY

In any video, two types of information present are spatial and temporal. The spatial information is the video frame's facial appearance, and the video frame movement in the video is the temporal information. The temporal information can be understood as the changes in facial gestures like the movement of lips, changes in eye sizes, etc. The example of recognition is like opening eyes as big and mouth too, which represents surprise. Here we have used this spatial and temporal information and built a framework that is trained as end-to-end for FER.

To get the FER from videos, we have designed the structure shown in Figure 1. We have used here 30 temporal and spatial CNN streams. Both streams aim to get the features from the videos, and then these features from the complete video are aggregated in the EmotionalVLAN layer of the CNN model. The output from the EmotionalVLAN is then forwarded to the Softmax layer as an input for the FER from videos. The training stage can give the most efficient parameters.

The CNN network builds for temporal signals is to study the temporal signals thoroughly. This network is fed with the optical flow displacement present in between two different video frames. Our work makes it easy to get FER efficiently as our framework need not estimate the motions of face implicitly. So the CNN network for temporal signals is designed based on this concept.

In next, we designed the static CNN network for processing the static video frames. It is helpful in FER that the static appearance is the most critical factor for getting facial expressions. We have used AlexNet here for implementing this spatial CNN network.

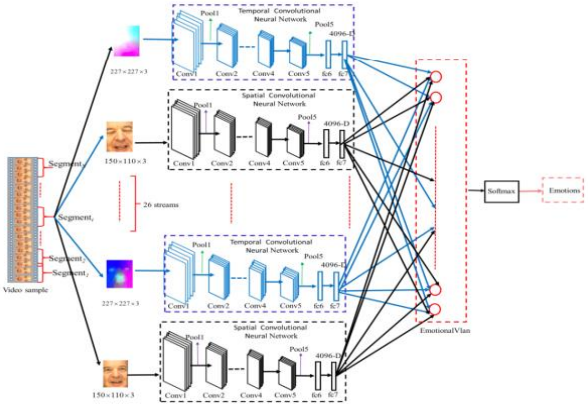

Figure 1. CNN Structure

As we have limited emotional video data, we designed the temporal and spatial CNN network based on the AlexNet architecture. We have convolution layers 5 in number; the max-pooling layer is 3 and fully connected layers is f6, f7, f8, f9. The max-pooling layer improves the overfitting problem. The f7 layer extracts the expressions from spatial and temporal

and gives it to the EmotionalVLAN for the generation of the features of the whole video. At last, the Softmax layer gets all the obtained features as input and classifies the facial expressions. 227*227*3 RGB images are provided as the input to the CNN. We initialize the AlexNet framework by giving some default values and then this framework is fine-tuned for the particular videos.

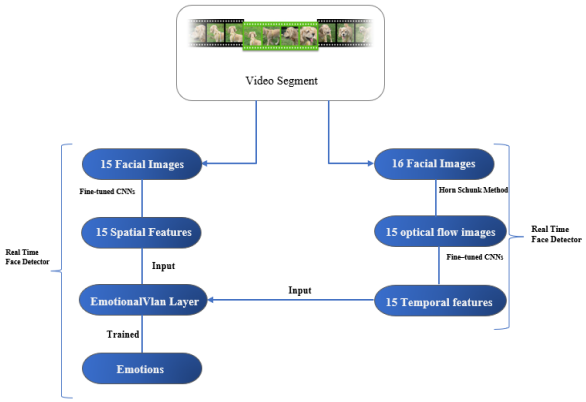

Figure 2. The flow of the proposed approach.

The flow of the proposed approach is shown in Figure 2. The steps involved in this work are stated below:

Step 1: The first step is to crop 16 facial images from

each video.

Step 2: Using these 16 facial images, we generated 15

optical flow images.

Step 3: We feed 15 spatial CNNs with 15 spatial facial

images, and 15 temporal CNNs with 15 optical

flow images.

Step 4: 15-15 spatial and temporal features were

extracted from fine-tuned spatial and temporal DCCNs.

Step 5: The Emotional VLAN receives 15–15 spatial

and temporal features for training and FER.

Algorithm for the proposed work is stated below:

Input: 15-15 static frames from the video and optical flow images

Output: Recognized facial expressions

1. Initialize AlexNet framework by 40 as mini-batch and rate of learning as 0.001
2. While epoch no<= 5000

// forward propagation
For every k neuron before the EmotionalVLAN layer {
F(n) = max (0, x) //output from the k neuron, x is input to the neuron

}

For every k neuron inside the EmotionalVLAN {
V[n] = $\sum_{i=1}^{T} e^{a_n} (x_{s,i} + x_{t.i})$ // $x_{t.i}$is the output from f7 layer for the temporal stream, $x_{s,i}$is the output from f7 layer for the spatial stream.
}

//Backpropagation error
For every neuron n in output layer {
$E_n = O_n(1 - O_n)(T_n - O_n)$ // error computation, Tn is the label for real emotion, ok is the label for the motional prediction
}
For each hidden neuron {
Weight update
}
}

Below we have given the three application stages: feature aggregation, video pre-processing, and the network's training.

a. Temporal-spatial layer of aggregation

Here in the proposed methodology, we have two representations of the spatial and temporal descriptor. $x_{s,i} \in R^D$ and $x_{t,i} \in R^D$ both here are used to present the D-dimensional spatial and temporal descriptors extracted from the i-th index of the video frame taken. In contrast, the temporal descriptors have it from i-th and i+1 th index. If seen technically, it is more convenient to overall aggregate video to get the $x_{s,i}$ and $x_{t,i}$ descriptors and save the overall content found in the video. And to do so, we need to assign every video descriptor with one of the available 4096 cells of EmotionalVLAN.

$$V[k] = \sum_{i=1}^{T} e^{a_k}(x_{s,i} + x_{t,i})$$

For the kth cell, we have a tunable parameter a_k. For an i^{th} frame of video, we have that addition of the $x_{s,i} + x_{t,i}$ as the temporal and spatial aggregation for the whole video. V[k] states the kth cell aggregation descriptor also we have v[k] = R ^4096*15 as one descriptor for video.

b. Video Preprocessing

Every video has various durations. Our input to the used framework has split our video into 16 frames because the interval we took for the

integer L is [2,20]. From this range, the best results have obtained on the value of L=16. Figure 2 shows the various L values comparisons in term of accuracy. In case any video has 16 plus segments, we drop the sections as (L-16)/2. In case of fewer parts than 16, we repeat section as (16-L)/2. When we have exact 16 parts, then the temporal images found are 15, as one temporal model comes from two video frames and reflects the change in between the 2 frames of video. We obtain the temporal information set of the motion vectors dt calculated from the two regular video frames, i.e., t, t+1. In the temporal image I_t we have dtx and dty as the change in two corresponding change in frames of video as t and t+1for horizontal and vertical respectively, and for this, we have used the Horn-Schunck method. We can compute the third channel to our temporal image as dtz. We have fed 15-15 static video frame and mundane pictures for the proposed approach.

$$d_t z = \sqrt{d_{t^2} x + d_{t^2} y}$$

We are cropping the face image by the real-time face detector to pre-process the spatial stream from the given frame of video. We have twice the width and thrice height from a distance between the two eyes in the facial image. So the vision we have cropped here is of size 150*110*3 as per the gap between both eyes.

c. Network Training

Here for our proposed approach, we trained our framework by the cross-entropy loss method, which is stated in figure1, this is the forwarding layer to the EmotionalVLAN. The Softmax finds the probability of the system. Here we are putting the EmotionalVLAN layer with the fc8 layer of AlexNet for every stream. After this, we have attached a new fc layer for the various categories of detected emotions in our system. We use the standard backpropagation method for minimizing the training stage, as shown below.
As shown in figure 1 we have 5 convolutional layers, 4 fc layers (fc6, fc7, fc8, fc9) and 3 max-pooling layers. Where EmotionalVLAN is the fc8 layer, and the fc6 and fc7 layers have 4096 units. And the fc9 is the layer for the various emotional category. Here in the model's training, we put both the optical and static images of 227*227*3.

Here the data available is X= { (bi, yi) } i=1,2,…N, here i is the number for a video frame, EmotionalVLAN layer's output is the bi, and the yi is the label for every segment in the video. The backpropagation method can be obtained by the below function for our network B:

$$min_{W^B \lambda^B} \sum_{i=1}^{N} H(softmax\left(W^B . \vartheta^B\left(b_i, \beta^b\right)\right), y_i)$$

Softmax layer weight is defined by W^B , $\vartheta^B\left(b_i, \beta^b\right)$ represents 122880-D ($R^{4096*30}$) as the output obtained from the EmotionalVLAN for the λ^B parameters. The given equation founds the log loss for the softmax layer:

$$H(B, y) = -\sum_{j=1}^{k} y_i \log({y_j}^B)$$

Total number of emotions is k, and the jth output for the softmax layer is the ${y_j}^B$.

3.1 Psedocode of Proposed Novel Algorithm

Step 1: From every video, we have cropped 16 facial images

Step 2: From this 16 facial images, we have generated 15 optical flow images.

Step 3: In 15 spatial CNNs, we are feeding 15 spatial facial images, and in the 15 temporal CNNs, we are supplying 15 optical flow images for fine-tuning.

Step 4: We have extracted 15-15 spatial and temporal features from fine-tuned spatial and temporal DCCNs.

Step 5: At last, the EmotionalVLAN gets this 15-15 spatial and temporal features for the training and FER.

Begin ()

{

V →FM *// cropping facial images from video*

FM→OM *// generating optical flow images from facial images*

*// fine-tuning OM's using CNN (i.e."**novel trainableGeneralized-Mean**") or other similar to it.*

AFT → F *// extracting features after fine-tuning using CNN*

Training (F); *// training model using spatial and temporal features*

Exit ();

}

Notations:

V – Videos

FM - fiscal images

OM – optical image

AFT – after fine-tuning

F – Features

IV. EXPERIMENT AND RESULTS ANALYSIS

We have conducted our experiments for the proposed approach on five datasets RML, BAUM-1s, eNTERFACE05, MMI, and FER2013. The implementation is carried out in the following manner. The CNN training is done with the values 40 for mini-batch and on .001 learning rate is taken. 1000 is the epoch number. MatconvNet is used for the implementation of CNN. For the training of the framework, the NVIDIA GPU of 25GB memory space is used. The subject independent cross-validation does testing of the framework performance. The datasets carry five subjects for evaluation. We have calculated the average accuracy of the framework for the utilization.

4.1 DATASETS:

PROPERTIES	BAUM 1s [17]	RML [18]	FER 2013 [20]	MMI [19]
Video samples	1222	720	35685	2900
Expressions	Joy, anger, sadness, disgust, fear, surprise, boredom, contempt	Anger, disgust, fear, Happiness, sadness, surprise	Anger, disgust, fear., neutral, Happiness, sadness, surprise	Anger, disgust, fear, joy, sadness, surprise
The frame size of the video	720*576*3		48*48 greyscale images	
People in video	31	8		75

Datasets we have used here is stated in the above table. Here we have cropped the images, and a selected number of videos has been taken for evaluation. The samples of all the datasets are shown below.

IMAGES OF DATASETS

Figure 3. Cropped facial images from BAUM-1s dataset [17]

Figure 4. Cropped facial images from MMI dataset [19]

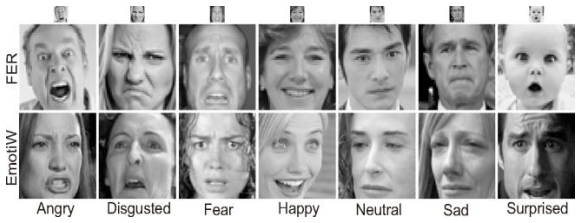

Figure 5. Cropped facial images from FER 2013 dataset [20]

Figure 6. Cropped facial images from RML dataset [18]

4.2 Experimental Result and Analysis:

Next is the evaluation of the temporal and spatial features that are learned earlier. Various streams in our framework are evaluated, and the accuracy is stated in table 3 for multiple datasets. Recognition from the spatial stream has 15 spatial streams, temporal streams have 15 temporal streams, and the spatial, temporal stream recognition involves 15 spatial and 15 temporal streams. Table 1 shows the accuracy obtained for the various streams and various datasets.

Features	**RML**	**BAUM-1s**	**FER-2013**	**MMI**
Spatial CNN	58.69	66.78	68.41	70.58
Temporal CNN	48.21	54.54	56.39	59.78

Table 1. Various Streams of Recognition Performance

The confusion matrix is also presented to understand the evaluation of facial expressions. The figures state that the accuracy obtained by the proposed approach for various datasets is higher, and the illustrations are evaluated with accuracy. It can be started from the confusion matrix that the fear of emotion is difficult to recognize as it comes to other facial expressions

Expression	Anger	Happy	Sad	Surprise	Neutral	Disgust
Anger	**79.0%**	0	7.5	0	0	2.5
Happy	0	**90**	0	0	5	0
Sad	0	0	**71.5**	2.5	5	0
Surprise	0	0	0	**89**	5	2.5
Neutral	0	0	0	7.5	**84.5**	0
Disgust	0	0	5	0	2.5	**78.5**
Average	**82.03**					

Table 2. Confusion Matrix for facial recognition on BAUM-1s dataset

Expression	Anger	Happy	Sad	Surprise	Neutral	Disgust
Anger	**74.00%**	0	0	0	2.5	2.5
Happy	0	**91.5**	0	0	2.5	0
Sad	0	0	**79.5**	0	2.5	2.5
Surprise	2.5	0	0	**88.5**	0	0
Neutral	0	0	0	0	**90**	5
Disgust	0	0	0	0	5	**83**
Average	**84.42**					

Table 3. Confusion Matrix for facial recognition on FER-2013 dataset

Expression	Anger	Happy	Sad	Surprise	Neutral	Disgust
Anger	**95.00%**	0	0	0	2.5	2.5
Happy	0	**97.5**	0	0	0	2.5
Sad	0	0	**97.5**	0	2.5	2.5
Surprise	2.5	0	0	**97.5**	0	0
Neutral	0	0	2.5	0	**95**	2.5
Disgust	5	0	0	0	0	**95**
Average	**96.25**					

Table 4. Confusion Matrix for facial recognition on MMI dataset.

Expression	Anger	Happy	Sad	Surprise	Neutral	Disgust
Anger	**90.00%**	0	10	0	0	0
Happy	2.5	**90**	0	0	7.5	0
Sad	0	10	**87.5**	2.5	0	0
Surprise	0	0	0	**92.5**	7.5	0
Neutral	0	5	5	0	**92.5**	0
Disgust	0	2.5	7.5	0	0	**90**
Average	**90.41**					

Table 5. Confusion Matrix for facial recognition on RML dataset

4.3 Comparison with different structures of DCNN:

We have compared the proposed framework with the various versions of DCNN methods for all the datasets used in our work. The table below states the values obtained for all purposes. It can be seen that the DCNN3 approach out performs the other two applied DCNN methods and can be used for future FER systems from videos.

DCNN Structure	**RML**	**BAUM-1s**	**FER-2013**	**MMI**
DCNN-1	59.35	68.42	73.51	69.78
DCNN-2	63.95	72.85	76.11	75.31
DCNN-3	69.82	76.49	81.39	80.49

Table 6. Comparison with state-of-the-art methods.

V. CONCLUSION

Our paper presents a novel approach to Facial Expression Recognition from the videos available. This work's proposed method aggregates the spatial and temporal features to classify the expressions present in videos for the gap current between the descriptors and the emotions seen in videos. The proposed method aggregates both convolutional features that are spatial and temporal present in the whole video. The proposed approach overcomes the limitations and performances of the previously applied state-of-the-art methods. The proposed plan also handles the overfitting problem. The presented system is tested on multiple datasets that are BAUM1s, FER 2013, MMI and RML.

Reference:

[1] S. Wang, S. Member, B. Pan, H. Chen, and Q. Ji, "Thermal Augmented Expression Recognition," pp. 1–12, 2018.

[2] X. Pan, S. Zhang, W. Guo, X. Zhao, Y. Chen, and H. Zhang, "Video-Based Facial Expression Recognition using Deep Temporal – Spatial Networks Video-Based Facial Expression Recognition using Deep Temporal – Spatial," vol. 4602, 2019, doi: 10.1080/02564602.2019.1645620.

[3] T. Ben, A. Radhouane, and M. Hammami, "Facial-expression recognition based on a low-dimensional temporal feature space," 2017.

[4] L. Greche, M. Akil, R. Kachouri, and N. Es, "ORIGINAL RESEARCH PAPER A new pipeline for the recognition of universal expressions of multiple faces in a video sequence,"*J. Real-Time Image Process.*, no. 0123456789, 2019, doi: 10.1007/s11554-019-00896-5.

[5] M. M. Hassan, M. Alrubaian, G. Fortino, and S. Member, "Facial Expression Recognition Utilizing Local Direction - based Robust Features and Deep Belief Network," vol. 3536, no. c, 2017, doi: 10.1109/ACCESS.2017.2676238.

[6] S. Lee, S. H. Lee, K. N. K. Plataniotis, and Y. M. Ro, "Experimental investigation of facial expressions associated with visual discomfort : Feasibility study towards an objective measurement of visual discomfort based on facial expression," no. c, 2016, doi: 10.1109/JDT.2016.2616419.

[7] E. S. Networks, K. Zhang, Y. Huang, Y. Du, and S. Member, "Facial Expression Recognition Based on Deep," vol. 14, no. 8, 2017, doi: 10.1109/TIP.2017.2689999.

[8] I. Latin and A. Transactions, "Deep Neural Network Architecture : Application for Facial Expression Recognition," vol. 18, no. 7, pp. 1311–1319, 2020.

[9] S. Agarwal, B. Santra, and D. Prasad, "Anubhav : recognizing emotions through facial expression,"*Vis. Comput.*, 2016, doi: 10.1007/s00371-016-1323-z.

[10] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial Expression Recognition in Video with Multiple Feature Fusion," vol. 3045, no. c, pp. 1–13, 2016, doi: 10.1109/TAFFC.2016.2593719.

[11] Y. Ding, Q. Zhao, B. Li, and X. Yuan, "Facial Expression Recognition from Image Sequence based on LBP and Taylor Expansion," vol. 3536, no. c, pp. 1–10, 2017, doi: 10.1109/ACCESS.2017.2737821.

[12] H. Kabir, S. Salekin, and Z. Uddin, "Facial Expression Recognition from Depth Video with Patterns of Oriented Motion Flow," vol. 3536, no. c, pp. 1–9, 2017, doi: 10.1109/ACCESS.2017.2704087.

[13] K. S. Yadav, "Facial expression recognition using modified Viola-John' s algorithm and KNN classifier," 2020.

[14] S. Liu, Y. Zhang, X. Yang, D. Shi, and J. J. Zhang, "Robust facial landmark detection and tracking across poses and expressions for in-the-wild monocular video," 2016, doi: 10.1007/s41095-016-0068-y.

[15] P. I. Rani and K. Muneeswaran, "Recognize the facial emotion in video sequences using eye and mouth temporal Gabor features,"*Multimed. Tools Appl.*, 2016, doi: 10.1007/s11042-016-3592-y.

[16] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning,"*IEEE Access*, vol. 7, pp. 32297–32304, 2019, doi: 10.1109/ACCESS.2019.2901521.

[17] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," IEEE Trans. Affective Comput., vol. 8, no. 3, pp. 300–313, Jul./Sep. 2016.

[18] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition," IEEE Trans. Multimedia, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[19] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Amsterdam, The Netherlands, Jul. 2005, pp. 317–321.

[20] https://www.kaggle.com/msambare/fer2013

[21] http://web3.bilkent.edu.tr/enterface19/call-for-projects/