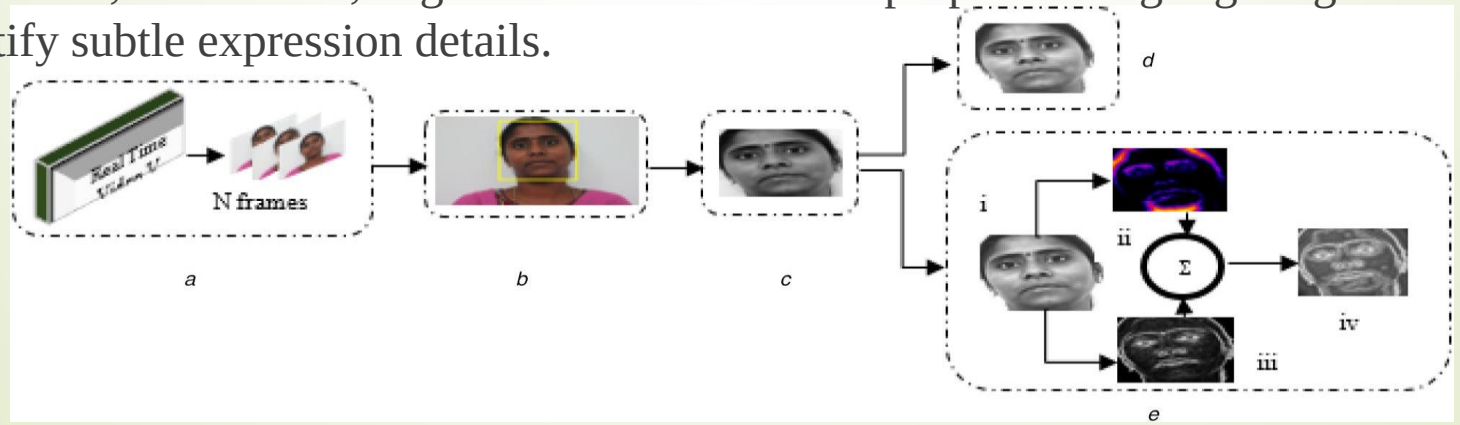
Facial Emotion Recognition with CNN and LSTM

- Aim:

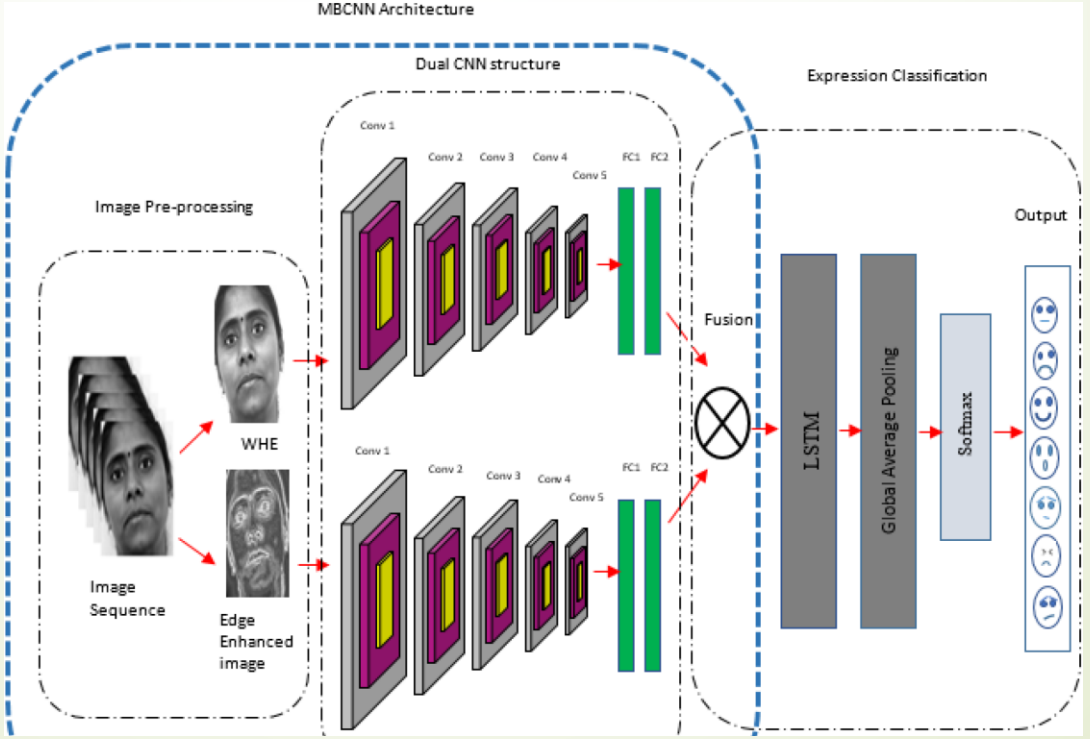
To carry out the task of FER with a novel deep learning model comprising of Maximum Boosted CNN and LSTM.

- Methodology:

  - Image pre-processing is performed on each frame of the input video sequence. The Viola-Jones algo. is first used for face detection followed by cropping and resizing the face to a 96 x 96 image size which is then converted to grayscale.
  - Two different image processing techniques are then performed on each image. First, a weighted histogram equalization (WHE) for the purpose of handling distinct illumination conditions, and second, edge enhancement for the purpose of highlighting latent features to identify subtle expression details.



- Both the WHE and Edge Enhanced Images are then passed to the same CNN structure with 5 convolutional, max pooling, batch norm and ReLU layers. This dual CNN structure along with the pre-processing step is know as Maximum Boosted CNN or MBCNN structure.
- Since this paper aims to target video datasets, focus needs to be laid on both the spatial and temporal aspects. Therefore, the CNN is used for feature extraction and to handle spatial signals while LSTMs are used for dealing with the temporal part of the problem.
- The outputs from the dual CNNs are fused using matrix addition and fed to the LSTM. This LSTM layer then generates a new representative feature vector. Finally, the softmax function performs classification.



- Results:
  - 4 datasets are used for the purpose of evaluating the proposed model: CK+, MMI, SFEW, Own dataset. Results of 2 of which are given below:

Table 2 Confusion matrix of CK+ database for six classes of expressions including neutral expression

	An	Di	Fe	Ha	Sa	Su	Ne
An	**96.29**	2.69	0	0	0	0	1.02
Di	1.57	**97.63**	0.8	0	0	0	0
Fe	0	0	**90.21**	4.17	0.22	5.4	0
Ha	0	0	0	**100**	0	0	0
Sa	0	0	0	0	**93.14**	0	6.86
Su	0	0	0.76	0	0	**98.02**	1.22
Ne	0.22	0	0	0	0.72	0	**99.06**

Table 3 Confusion matrix of the MMI database for six classes of expressions including neutral expression

	An	Di	Fe	Ha	Sa	Su	Ne
An	**68.32**	9.31	6.2	0	7.29	0	8.88
Di	7.26	**81.22**	0	5.07	4.53	0	1.92
Fe	2.09	0	**77.64**	0	0	12.16	8.11
Ha	0	3.62	0	**93.28**	0	0	3.1
Sa	6.12	3.56	0	0	**71.24**	0	19.08
Su	0	0	9.26	0	0	**83.43**	7.31
Ne	5.36	0	2.54	2.01	3.02	0	**87.06**

  - The proposed model, in general performs better on happiness, surprise and neutral expressions and worse off on fear, disgust and angry expressions.

    The model's performance is comparable to existing state-of-the-art models on these datasets.

Table 7 Overall performance comparisons on CK+, MMI and SFEW datasets with the state-of-the-art methods

References/year	Methods	Database accuracy			
		CK+, %	MMI, %	SFEW, %	OWN, %
Liu *et al.*/2014 [53]	3DCNN-DAP	92.40	63.40	—	—
Jung *et al.*/2015 [38]	DTAGN (joint)	97.25	70.24	—	—
Liang *et al.*/2019 [54]	DSN + DTN + BiLSTM	99.60	80.71	—	—
Salmam *et al.*/2019 [55]	CNN–DNN	96.92	—	—	—
Zhang *et al.*/2017 [56]	MSCNN-PHRNN	98.50	81.18	—	—
Mollahosseini *et al.*/2016 [35]	DNN	93.20	77.90	47.70	—
Yu and Zhang/2018 [57]	ensemble of deep CNN	—	—	61.29	—
Levi and Hassner/2015 [58]	mapped LBP	—	—	54.56	—
input 1 (illuminance corrected)	WHE + CNN + LSTM	88.02	61.12	32.33	85.05
input 2 (edge enhanced)	edge enhancement + CNN + LSTM	96.33	74.0	43.21	92.45
input 1 and 2	MBCNN	96.21	78.45	47.36	93.00
proposed method	MBCNN + LSTM	99.01	81.60	56.68	95.21