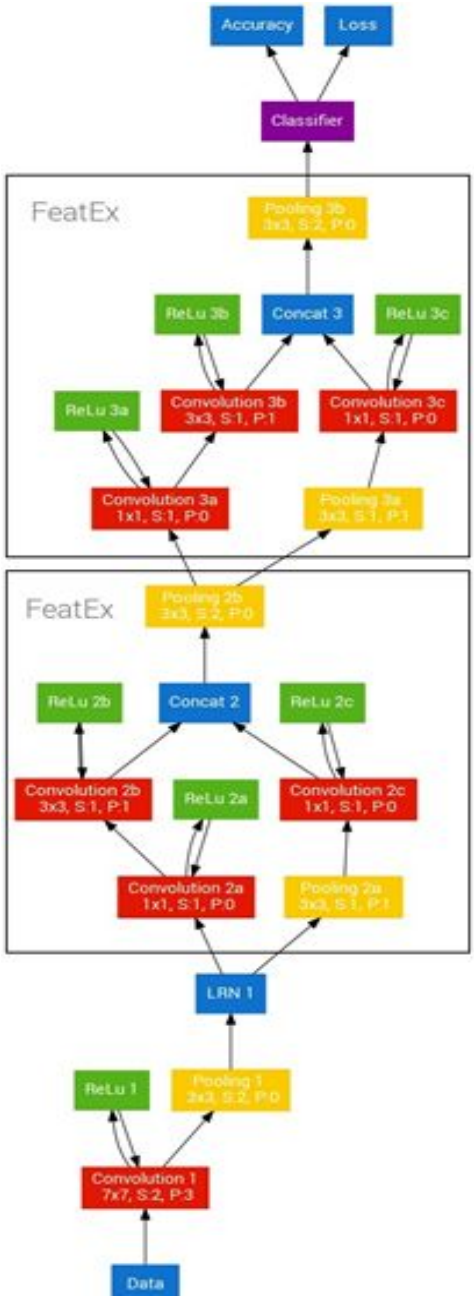
EMOTION RECOGNITION USING FACIAL EXPRESSIONS FOR IMAGES

Introduction

Methods for emotion recognition often involve the *Facial Action Coding System (FACS)* which describes the facial expression using *Action Units (AU)*. An Action Unit is a facial action like ”raising the Inner Brow”. Detecting such landmarks can be hard, as the distance between them differs depending on the person .
The presented approach uses *Convolutional Neural Networks* special type of *Artificial Neural Networks(ANNs)*.The proposed network has been trained on the **FER-2103 Dataset** and evaluated on the **CK+ dataset**.

Model Overview



Layer	Output Size
Data	48x48x1
Convolutional 1	21x21x64
Pooling 1	10x10x64
Batch Normalization(LRN)	10x10x64
Convolutional 2a	10x10x96
Convolutional 2b	8x8x208
Pooling 2a	8x8x64
Convolutional 2c	8x8x64
Concat 2	8x8x272
Pooling 2b	6x6x272
Convolutional 3a	8x8x96
Convolutional 3b	6x6x208
Pooling 3a	6x6x272
Convolutional 3c	6x6x64
Concat 3	6x6x272
Flatten 1	9792
Dropout 1	9792
Dense 1	7

Pre Processing

- **Resizing**: Initially the image size is *(48,48)*. The image is resized to *(480,480)* for the *Viola Jones Algorithm* to work. This facilitates better face and eye detection.
- **Face Detection**: The second step involves face detection. Face is detected using *The Viola Jones face detection algorithm*.
- **Rotation**: Landmark of Eyes is used to align the face with the horizontal. Using Viola Jones Detection Algorithm, eyes are detected, the centroid of each eye is calculated. Calculate the angle between the horizontaal and the line joining the centroid of eyes. *Transformation Matrix* is calculated using opencv application and then final rotation is done.
- **Cropping:** The distance between eyes is used as a parameter for cropping. Detect the eyes of the face using the Viola-Jones algorithm, calculate the centroid for each eye and then calculate the distance between them. The width and height for cropping was set as follows:
  *The width of image = 2.1 * distance between eyes.*

*The height of image = 3.2 * distance between eyes.*

- **Smoothing:** Smoothing is done to remove noise from the image. *Bilateral smoothing* is used to preserve the edges in the face.
- **Resizing:** Finally after all the preprocessing, images are resized to *(48,48)* which is fed into the model.

Proposed Architecture

- The proposed model was trained on the *FER-2013* dataset for *13* epochs with a batch-size of *128* images.
- The two *FeatEx* (Parallel Feature Extraction Block) blocks inspired by the success of *GoogleNet* consist of Convolutional, Pooling, and ReLU Layers. The first Convolutional layer in FeatEx reduces the dimension since it convolves with a filter of size 1 × 1. It is enhanced by a ReLU layer, which creates the desired sparseness. The output is then convolved with a filter of size 3 × 3. In the parallel path a Max Pooling layer is used to reduce information before applying a CNN of size 1 × 1. The paths are concatenated for a more diverse representation of the input. Using this block twice yields good results.
- We used *K-cross Validation technique* to improve generalization accuracy and prevent overfitting. For *10 fold cross validation* the model took nearly 5 hours of training

Experiments and Results

FER-2013: For the experiments all *35887* annotated images of size 48x48 have been used to do a 10-fold cross- validation. The proposed architecture has proven to be very effective on this dataset with an average accuracy of **95.6%**.The max accuracy proposed by paper was **98%.**

The accuracy on the FER-2013 set shows that the chosen approach is robust, misclassification usually occurs on pictures which are the first few instances of an emotion sequence.

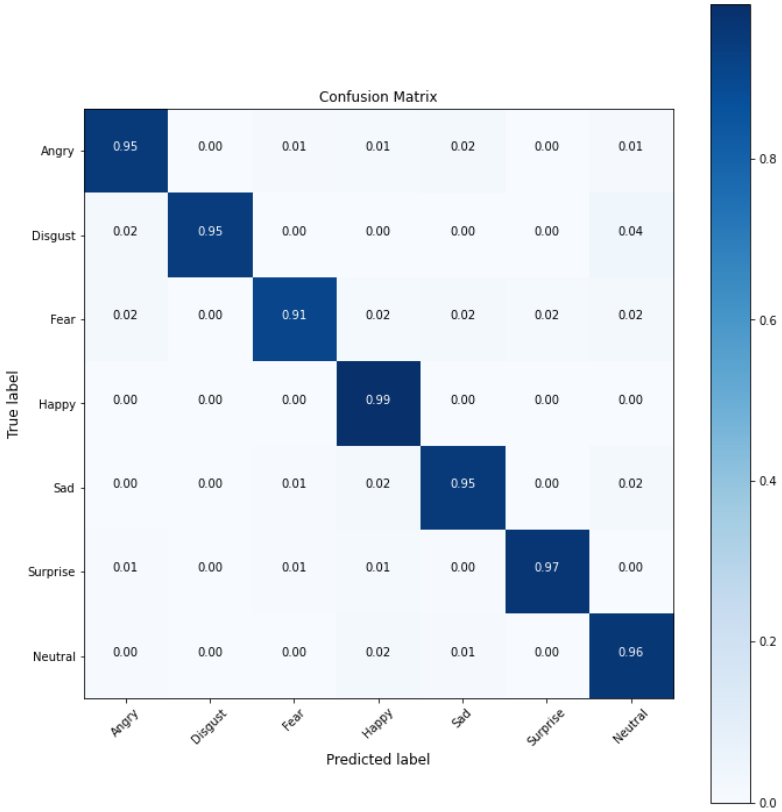
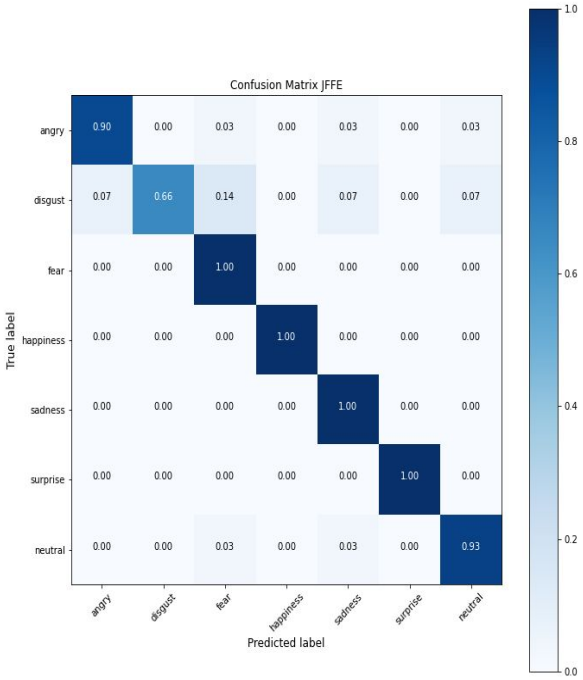
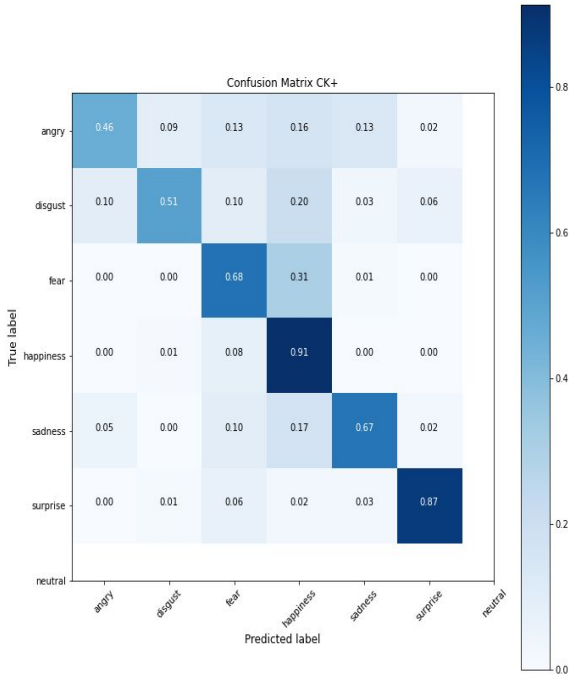


Table 1: Accuracies of models

Model	Database	Accuracy
Deep Emotion(proposed)	JAFFE	~92.8 %
Deep Emotion(proposed)	CK+	~89.3 %
Our implementation	JAFFE	~97 %
Our implementation	CK+	~59 %



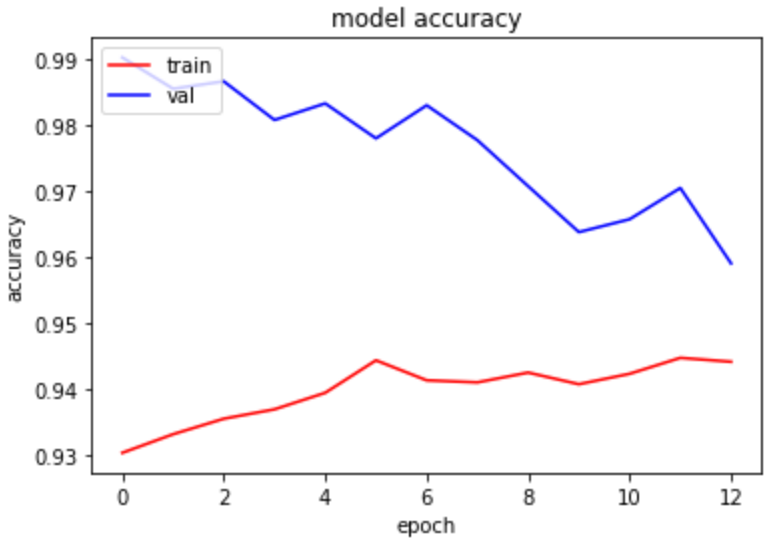


Average accuracy on **JAFFE** dataset ~ *97%*.
The lowest accuracy is achieved by *Disgust* with *66%* while *Happy, Fear, Sad* and *Surprise* are recognized *100%*.

Accuracy on **CK+** dataset ~ *59%*.
The lowest accuracy is achieved by *Angry* with *46%* while *Happy* is recognized with *91%*.

Visualisation and Validation

- Using confusion matrix to analyse which emotions are lacking accuracy and working on that to improve upon them
- Also using the graph for validation accuracy and training accuracy to check if the model doesn't overfit.
- We got maximum accuracy of 99%(Happy) and minimum accuracy of 91%(Fear).
- We observed that the model is able to predict Happiness extremely well but gets confused with Fear,Disgust and Angry.



References

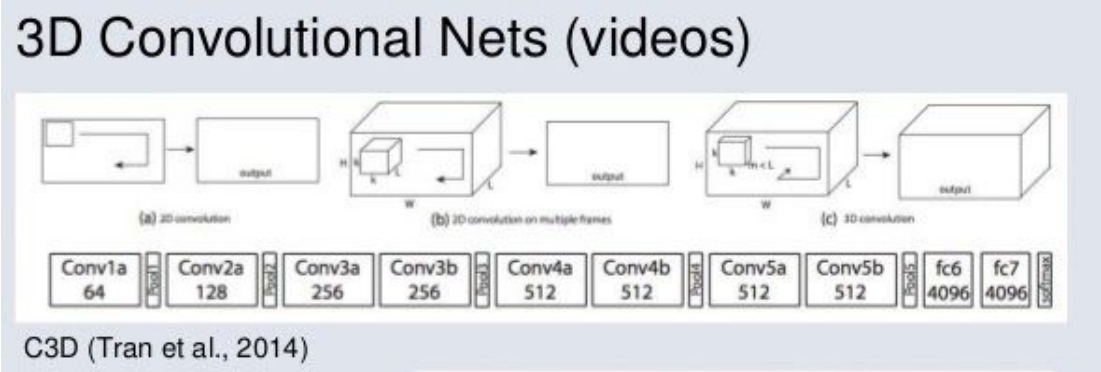
1. DeXpression: Deep Convolutional Neural Network for Expression Recognition
2. Emotion-recognition: Real time emotion recognition(For real-time prediction using webcam)
3. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network(https://arxiv.org/pdf/1902.01019.pdf?fbclid=IwAR0E0k87fcKawejhdtO5VvpZT24FVuuM3tlWoy_1x3e_JFW2dB6Z7wGAS3A)
4. https://stackoverflow.com

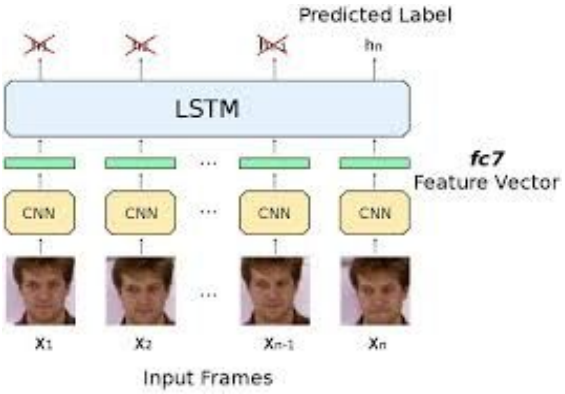
EMOTION RECOGNITION USING FACIAL EXPRESSIONS FOR VIDEOS

Introduction

Many researchers have tried to identify emotions in videos based on computer vision technologies.Traditional convolutional neural networks have a major limitation that they just handle spatial information and ignore the temporal information.
The paper proposes a hybrid of C3D and CNN+LSTM network. LSTM has memory ability and suits for processing sequences with contexts well.

Model Overview





Dataset:[BAUM-2]

Total number of videos: 1047
Frames per second: 23
Frames Extracted per clip: 16
Average duration of clips: 1-2 sec
No. of videos of each emotion:

1. Happy: 248
2. Sad: 137
3. Anger: 173
4. Disgust: 51
5. Fear: 68
6. Surprise: 152
7. Neutral: 169

Pre-processing

- From each video 12 frames are extracted for further preprocessing.
- For each frame we apply Viola Jones algorithm to detect face.
- Rotation: The face is rotated using landmarks of eyes as reference. The angle between horizontal line and line between eyes is used to get the transformation matrix and then the face is rotated.
- Cropping: The distance between eyes is used as reference for cropping.
- Furthermore we apply a bilateral filter for smoothing and resize each frame to 48x48.
- Video frames are aligned using similarity transform using facial key points.

Feature extraction+Classification

Approach-1(LSTM):

- Each frame of a video is fed to our static image model(CNN) fine tuned on BAUM-2 dataset.
- We remove the Softmax activation layer from the CNN to produce a feature vector for each frame of size 128.
- Then we pass the image embeddings through a single layer LSTM with 128 embedding outputs and pass the output of the last layer through Softmax layer to obtain the results.

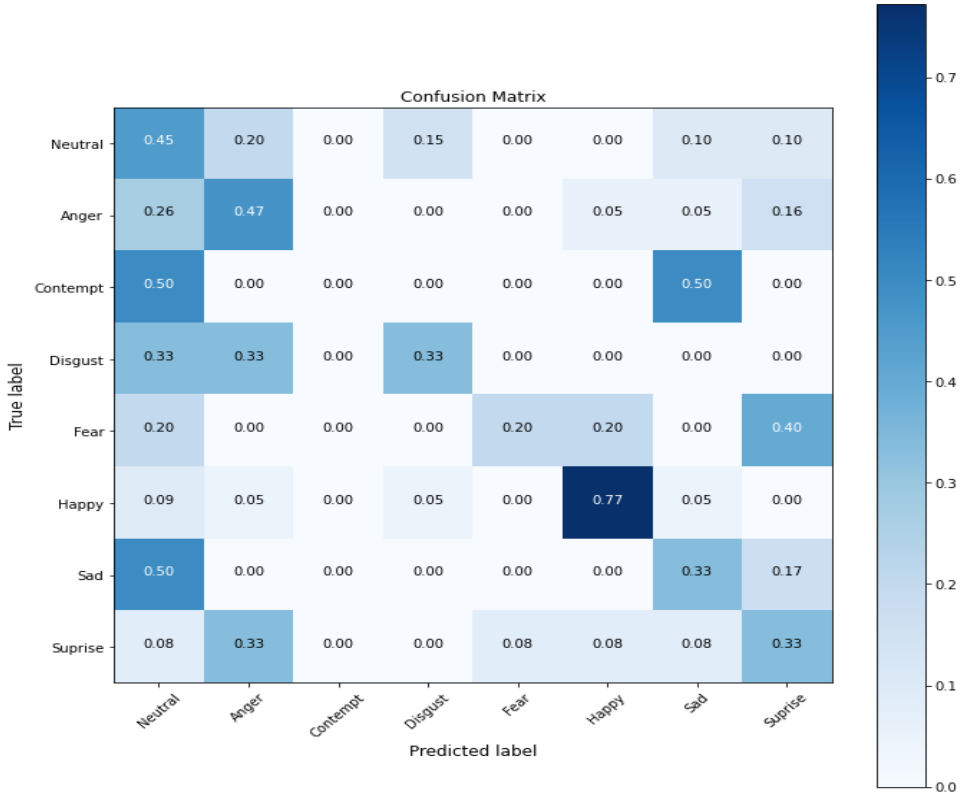
Approach-2(C3D):

- In this we build a 3D convolutional network from scratch and train it on BAUM-2 dataset
- In this we pass the input of shape 12x48x48x1 as one input where the model perceives 12 as depth of input and 48,48 as height and width of input with a single channel.
- After 8 Conv3D layers and 4 MaxPool3D layers we have added 2 fully connected layers and then a Softmax activation layer to obtain the results.

Results

Table 1: Accuracy of models

Model	Database	Accuracy
Single VGG+LSTM(proposed)	AFEW(7 emotions)	~39.6 %
Single C3D(proposed)	AFEW(7 emotions)	~38.7 %
1CNN+3C3Ds (proposed)	AFEW(7 emotions)	~59.2%
Our implementation of CNN+LSTM	BAUM(8 emotions)	~36 %
Our implementation of C3D	BAUM(8 emotions)	~28.4 %



:

References

1. Video-based emotion recognition using CNN-RNN and C3D hybrid networks
2. https://arxiv.org/pdf/1711.04598.pdf
3. https://github.com/TianzhongSong/C3D-keras/blob/master/models.py