Facial Emotion Recognition Using Convolutional Neural Network

N Rishmabegam[1], S I Padma[2]
[1]Dept of ECE
[2]Assistant Professor, Dept of ECE
[1, 2] PET Engineering College Tirunelveli

Abstract- *Facial expression recognition is challenging due to various un- constrained conditions. Here propose a convolutional neural networks (CNN) model can be trained to analyze images and identify face emotion. We create a system that recognizes students' emotions from their faces. CNN is an end to end learning framework. It combines the multiple representations from facial regions of interest (ROIs). Each representation is weighed via a proposed Gate Unit that computes an adaptive weight from the region itself according to the unobstructed-ness and importance. Our system consists of emotion recognition using CNN on FER 2013 database with seven types of expressions. Obtained results show that face emotion recognition is feasible in education; consequently, it can help teachers to modify their presentation according to the students' emotions.*

Keywords- human computer interface, facial expression recognition, convolutional neural network.

I. INTRODUCTION

Emotional expressions are very important in human communication. As interactive technology has become ubiquitous in our society and is taking on coaching roles in education, emotion recognition from facial expressions has become a crucial part in human-computer interaction. However, human facial expressions change so subtly that automatic facial expression recognition has always been a challenging task. Here a method has been used to achieve facial expression recognition using convolutional neural network. Facial expression recognition (FER) has received significant interest from computer scientists and psychologists over recent decades, as it holds promise to an abundance of applications, such as human-computer interaction, affect analysis, and mental health assessment. Although many facial expression recognition systems have been proposed and implemented, majority of them are built on images captured in controlled environment, such as CK+, MMI, Oulu-CASIA, and other lab-collected datasets. The controlled faces are frontal and without any occlusion. The FER systems that perform perfectly on the lab-collected datasets, are probable to perform poorly when recognizing human expressions under natural and un-controlled conditions.

Facial expression recognition has brought much attention in the past years due to its impact in clinical practice, sociable robotics and education. According to diverse research, emotion plays an important role in education. Currently, a teacher use exams, questionnaires and observations as sources of feedback but these classical methods often come with low efficiency. Using facial expression of students the teacher can adjust their strategy and their instructional materials to help foster learning of students. The purpose of this article is to implement emotion recognition in education by realizing an automatic system that analyze students' facial expressions based on Convolutional Neural Network (CNN), which is a deep learning algorithm that are widely used in images classification. It consist of a multistage image processing to extract feature representations. Our system includes three phases: face detection, normalization and emotion recognition that should be one of these seven emotions: neutral, anger, fear, sadness, happiness, surprise and disgust.

Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications. Facial expression analysis has been an active research topic for behavioral scientists since the work of Darwin in 1872. Suwa et al. presented an early attempt to automatically analyze facial expressions by tracking the motion of 20 identified spots on an image sequence in 1978. After that, much progress has been made to build computer systems to help us understand and use this natural form of human communication.

Facial expression analysis refers to computer systems that attempt to automatically analyze and recognize facial motions and facial feature changes from visual information. Sometimes the facial expression analysis has been confused with emotion analysis in the computer vision domain. For emotion analysis, higher level knowledge is required. For example, although facial expressions can convey emotion, they can also express intention, cognitive processes, physical

effort, or other intra- or interpersonal meanings. Interpretation is aided by context, body gesture, voice, individual differences, and cultural factors as well as by facial configuration and timing. Computer facial expression analysis systems need to analyze the facial actions regardless of context, culture, gender, and so on.

The accomplishments in the related areas such as psychological studies, human movement analysis, face detection, face tracking, and recognition make the automatic facial expression analysis possible. Automatic facial expression analysis can be applied in many areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human computer interface (HCI).

II. LITERATURE SURVEY

In the paper titled "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression" T. Kanade et.al (2010) proposed the Cohn-Kanade database in 2000 which was released for the purpose of promoting research into automatically detecting individual facial expressions. Since then, the CK database has become one of the most widely used test-beds for algorithm development and evaluation. During this period, three limitations have become apparent: 1) While AU codes are well validated, emotion labels are not, as they refer to what was requested rather than what was actually performed, 2) The lack of a common performance metric against which to evaluate new algorithms, and 3) Standard protocols for common databases have not emerged. As a consequence, the CK database has been used for both AU and emotion detection (even though labels for the latter have not been validated), comparison with benchmark algorithms is missing, and use of random subsets of the original database makes meta-analyses difficult. To address these and other concerns, we present the Extended Cohn-Kanade (CK+) database. The number of sequences is increased by 22% and the number of subjects by 27%. The target expression for each sequence is fully FACS coded and emotion labels have been revised and validated. In addition to this, non-posed sequences for several types of smiles and their associated metadata have been added. Baseline results are presented using Active Appearance Models (AAMs) and a linear support vector machine (SVM) classifier using a leave-one-out subject cross-validation for both AU and emotion detection for the posed data. The emotion and AU labels, along with the extended image data and tracked landmarks will be made available July 2010.

In the paper titled "Web-based database for facial expression analysis" M. Pantic et.al (2005) proposed that in the last decade, the research topic of automatic analysis of facial expressions has become a central topic in machine vision research. Nonetheless, there is a glaring lack of a comprehensive, readily accessible reference set of face images that could be used as a basis for benchmarks for efforts in the field. This lack of easily accessible, suitable, common testing resource forms the major impediment to comparing and extending the issues concerned with automatic facial expression analysis. In this paper, we discuss a number of issues that make the problem of creating a benchmark facial expression database difficult. We then present the MMI facial expression database, which includes more than 1500 samples of both static images and image sequences of faces in frontal and in profile view displaying various expressions of emotion, single and multiple facial muscle activation. It has been built as a Web-based direct-manipulation application, allowing easy access and easy search of the available images. This database represents the most comprehensive reference set of images for studies on facial expression analysis to date.

In the paper titled "Facial expression recognition from near-infrared videos" M. Pantic (2011) proposed the facial expressions can be thought as specific dynamic textures where local appearance and motion information need to be taken into account. Local spatiotemporal operators to describe facial expressions has been utilized. All current facial expression recognition databases are captured in visible light spectrum. Visual light usually changes with locations, and can also vary with time, which can cause significant variations in image appearance and texture. In this paper, we present a novel research on a dynamic facial expression recognition from near-infrared (NIR) video sequences. NIR imaging is robust with respect to illumination changes. Experiments on a new NIR database show promising and robust results against illumination variations.

In this paper titled "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild" W. Deng (2017) proposed that facial expression is central to human experience, but most previous databases and studies are limited to posed facial behavior under controlled conditions. In this paper, A novel facial expression database, Real-world Affective Face Database (RAF-DB), which contains approximately 30000 facial images with uncontrolled poses and illumination from thousands of individuals of diverse ages and races has been presented. During the crowdsourcing annotation, each image is independently labeled by approximately 40 annotators. An expectation-maximization algorithm is developed to reliably estimate the emotion labels, which reveals that real-world faces often

express compound or even mixture emotions. A cross-database study between RAF-DB and CK+ database further indicates that the action units of real-world emotions are much more diverse than, or even deviate from, those of laboratory-controlled emotions. To address the recognition of multi-modal expressions in the wild, we propose a new deep locality-preserving convolutional neural network (DLP-CNN) method that aims to enhance the discriminative power of deep features by preserving the locality closeness while maximizing the inter-class scatter. Benchmark experiments on 7-class basic expressions and 11-class compound expressions, as well as additional experiments on CK+, MMI, and SFEW 2.0 databases, show that the proposed DLP-CNN outperforms the state-of-the-art handcrafted features and deep learning-based methods for expression recognition in the wild. To promote further study, we have made the RAF database, benchmarks, and descriptor encodings publicly available to the research community.

III. PROPOSED SYSTEM

This project presents analyze students' facial expressions using a Convolutional Neural Network (CNN) architecture. First, the system detects the face from input image and these detected faces are cropped and normalized to a size of 48×48. Then, these face images are used as input to CNN. Finally, the output is the facial expression recognition results (anger, happiness, sadness, disgust, surprise or neutral). Figure 1 presents the structure of our proposed approach.

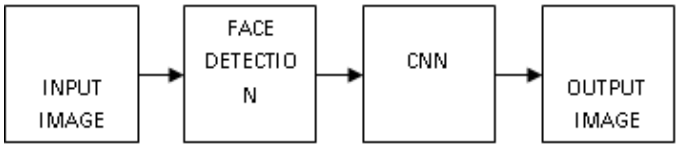


Fig 1 Block Diagram

The proposed convolutional neural network with attention mechanism (ACNN) for facial expression recognition with partial occlusions. To address the occlusion issue, ACNN endeavours to focus on different regions of the facial image and weighs each region according to its obstructed-ness (to what extent the patch is occluded) as well as its contribution to FER. This project is based on detecting 22 most important facial feature point and generation of facial feature vector by finding the Euclidian distances between some particular points. The key facial regions are found based on a defined face model. The face model is generated by the detection of the eyes and mouth points. Feed forward back Propagation neural network is used as the network classifier to classify the facial expression from a set of seven basic expressions like happy, sad, surprise, fear, anger, disgust and neutral. The experiment is done on Color FERET Database and got an accuracy of 100% for trained dataset and 85% accuracy for test set. The image is fed into a convolutional net (VGG) and is represented as some feature maps.

Then, ACNN decomposes the feature maps of the whole face to multiple sub-feature maps to obtain diverse local patches. Each local patch is encoded as a weighed vector by a Patch-Gated Unit (PG-Unit). A PG-Unit computes the weight of each patch by an Attention Net, considering its obstructed-ness. Besides the weighed local representations, the feature maps of the whole face are encoded as a weighed vector by a Global-Gated Unit (GG-Unit). The weighed global facial features with local representations are concatenated and serve as a representation of the occluded face. Two fully connected layers are followed to classify the face to one of the emotional categories. ACNNs are optimized by minimizing the softmax loss. Considering different interest of the local and global regions, we introduce two versions of ACNN: Patch based ACNN (pACNN) and Global-local based ACNN (gACNN). pACNN only contains local attention mechanism.

CONVOLUTIONAL NEURAL NETWORK (CNN)

A Convolutional Neural Network (CNN) is a deep artificial neural networks that can identify visual patterns from imput image with minimal pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. The important unit inside a CNN layers is a neuron. They are connected together, in order that the output of neurons at a layer becomes the input of neurons at the next layer.

In order to compute the partial derivatives of the cost function the backpropagation algorithm is used. The term convolution refers to the use of a filter or kernel on the input image to produce a feature map. In fact, CNN model contains 3 types of layers

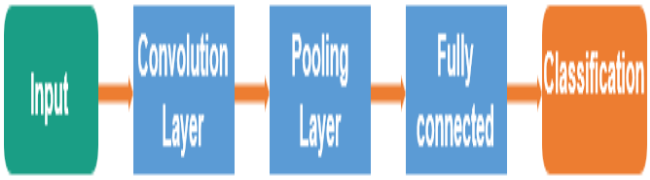


Fig 2 CNN architecture

Convolution Layer: is the first layer to extract features from an input image. The primary purpose of Convolution in case of a ConvNet is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data. It

performs a dot product between two matrices, where one is the image and the other is a kernal. The convolution formula is represented in Equation 1 :

$$net(t,f) = (x * w)[t,f] = \sum^{m}\sum^{n} x[m,n]w[t-m,f-n] \quad (1)$$

Where net(t, f) is the output in the next layer, x is the input image, w is the filter matrix and * is the convolution operation.

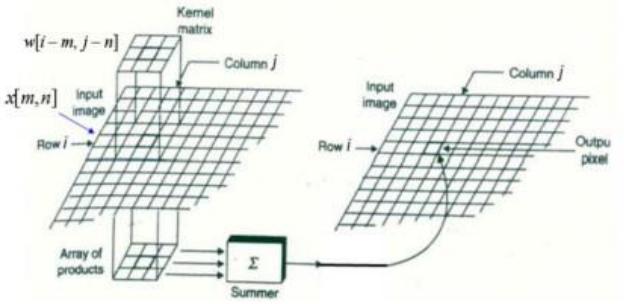


Fig 3 Details on Convolution layer

Pooling Layer: reduces the dimensionality of each feature map but retains the most important information. Pooling can be of different types : Max Pooling, Average Pooling and Sum Pooling. The function of Pooling is to progressively reduce the spatial size of the input representation and to make the network invariant to small transformations, distortions and translations in the input image. In our work, we took the maximum of the block as the single output to pooling layer.

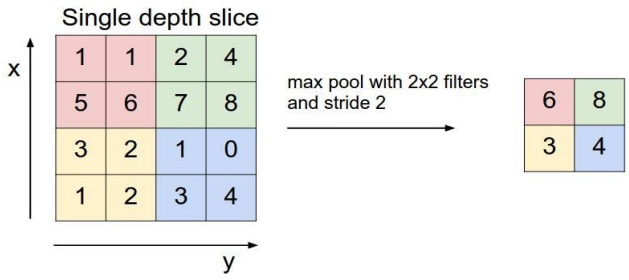


Fig 4 Details on Pooling layer

Fully connected layer: is a traditional Multi Layer Perceptron that uses an activation function in the output layer. The term "Fully Connected" implies that every neuron in the previous layer is connected to every neuron on the next layer. The purpose of the Fully Connected layer is to use the output of the convolutional and pooling layers for classifying the input image into various classes based on the training dataset. So the Convolution and Pooling layers act as Feature Extractors from the input image while Fully Connected layer acts as a classifier

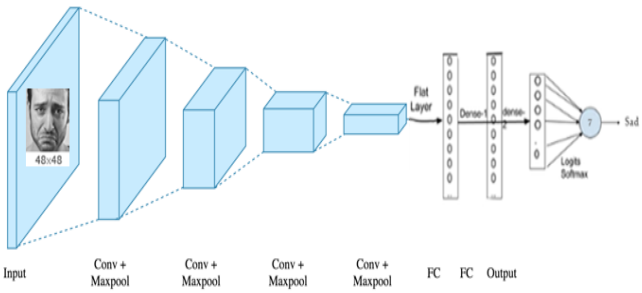


Fig 5 Details on Pooling layer

CNN model contains 4 convolutional layers with 4 pooling layers to extract features, and 2 fully connected layers then the softmax layer with 7 emotion classes. Input image is grayscale face image with a size of 48×48. For each convolutional layer we used 3×3 filters with stride 2. For the pooling layers, we used max pooling layer and 2×2 kernels with stride 2. Thus, to introduce the non linearity in our model we used the Rectified Linear Unit (ReLU), defined in Equation 2, wich is the most used activation function recently.

$$R(z) = \max(0, z)$$

R(*z*) is zero when *z* is less than zero and R(*z*) is equal to *z* when *z* is above or equal to zero. Table I presents the network configuration of our model.

FACIAL EXPRESSION RECOGNITION

A facial expression is one or more motions or positions of the muscles beneath the skin of the face. According to one set of controversial theories, these movements convey the emotional state of an individual to observers. Facial expressions are a form of nonverbal communication. They are a primary means of conveying social information between humans, but they also occur in most other mammals and some other animal species. The pioneer F-M Facial Action Coding System 2.0 was created in 2017 by Dr. Freitas-Magalhães, and presents about 2,000 segments in 4K, using 3D technology and automatic and real-time recognition.

Humans can adopt a facial expression voluntarily or involuntarily, and the neural mechanisms responsible for controlling the expression differ in each case. Voluntary facial expressions are often socially conditioned and follow a cortical route in the brain. Conversely, involuntary facial expressions are believed to be innate and follow a sub cortical route in the brain.

Facial recognition is often an emotional experience for the brain and the amygdala is highly involved in the recognition process.

The eyes are often viewed as important features of facial expressions. Aspects such as blinking rate can be used to indicate whether or not a person is nervous or whether or not he or she is lying. Also, eye contact is considered an important aspect of interpersonal communication. However, there are cultural differences regarding the social propriety of maintaining eye contact or not.

Beyond the accessory nature of facial expressions in spoken communication between people, they play a significant role in communication with sign language. Many phrases in sign language include facial expressions in the display.There is controversy surrounding the question of whether or not facial expressions are worldwide and universal displays among humans. Supporters of the Universality Hypothesis claim that many facial expressions are innate and have roots in evolutionary ancestors.

Opponents of this view question the accuracy of the studies used to test this claim and instead believe that facial expressions are conditioned and that people view and understand facial expressions in large part from the social situations around them.

FEATURE EXTRACTION

Feature extraction involves reducing the amount of resources required to describe a large set of data. General face features are forehead type and size, eye brows, eye color, nose shape and length. Mouth lips, chin shape and size, ear, hair and face shape micro expressions. Facial expressions of emotions are universal not learned differently in each culture.

Transforming the input data in to the set of features is called feature extraction. If the feature extracted are carefully chosen it is expected that the feature set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full-size input.

Extraction of discriminative features from salient facial patches plays a vital role in effective facial expression recognition. The accurate detection of facial landmarks improves the localization of the salient patches on face images. This paper proposes a novel framework for expression recognition by using appearance features of selected facial patches. A few prominent facial patches, depending on the position of facial landmarks, are extracted which are active during emotion elicitation. These active patches are further processed to obtain the salient patches which contain discriminative features for classification of each pair of expressions, thereby selecting different facial patches as salient for different pair of expression classes. One-against-one classification method is adopted using these features.

Facial expression, being a fundamental mode of communicating human emotions, finds its applications in human-computer interaction (HCI), health-care, surveillance, driver safety, deceit detection etc. Tremendous success being achieved in the fields of face detection and face recognition, affective computing has received substantial attention among the researchers in the domain of computer vision. Signals, which can be used for affect recognition, include facial expression, paralinguistic features of speech, body language, physiological signals (e.g. Electromyogram (EMG), Electrocardiogram (ECG), Electrooculogram (EOG), Electroencephalography (EEG), Functional Magnetic Resonance Imaging (fMRI) etc.). A review of signals and methods for effective computing, according to which, most of the research on facial expression analysis are based on detection of basic emotions: anger, fear, disgust, happiness, sadness, and surprise. A number of novel methodologies for facial expression recognition have been proposed over the last decade.

FEATURE SELECTION

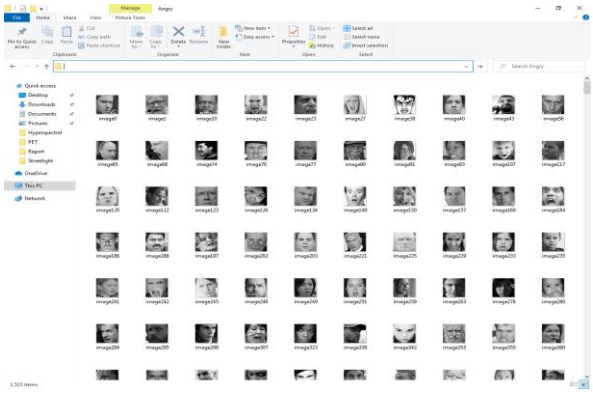
In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

- Simplification of models to make them easier to interpret by researchers/users,
- Shorter training times,
- To avoid the curse of dimensionality,
- Enhanced generalization by reducing overfitting (formally, reduction of variance)

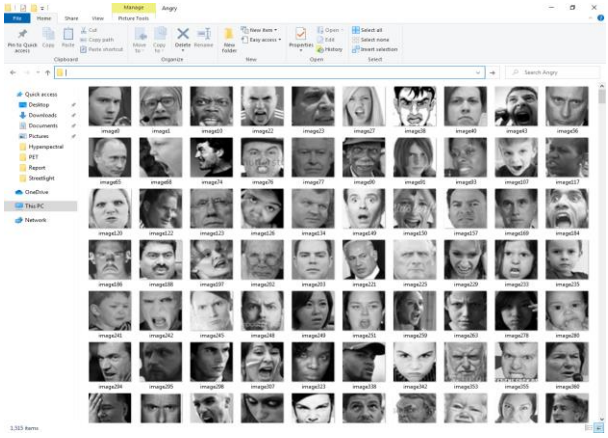
The central premise when using a feature selection technique is that the data contains many features that are either redundant or irrelevant and can thus be removed without incurring much loss of information. Redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated. Feature selection techniques should be distinguished from feature extraction. Feature extraction creates new features from functions of the

original features, whereas feature selection returns a subset of the features.

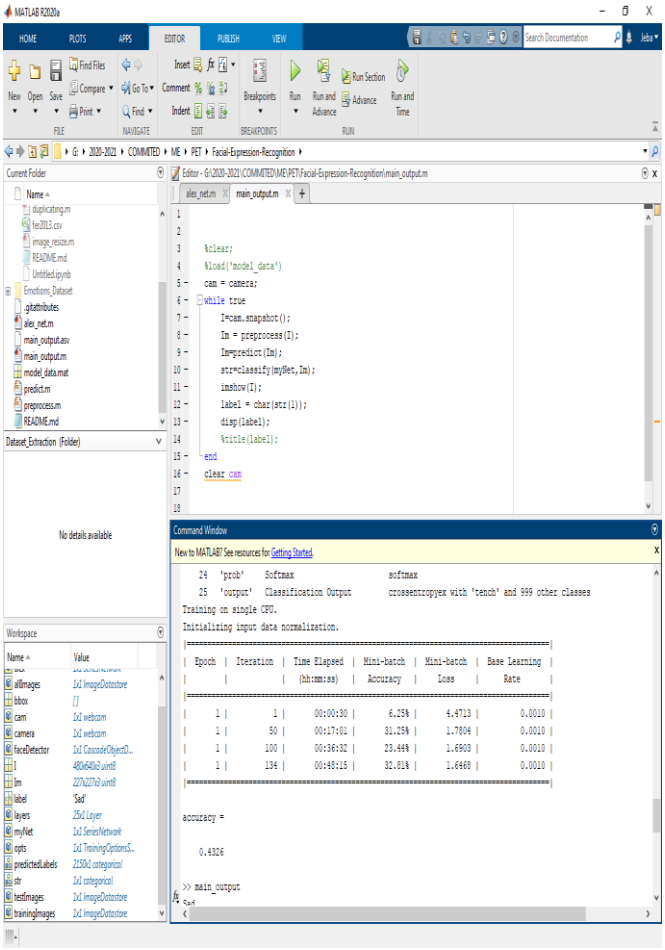
IV. SCREEN SHOTS



Input Image

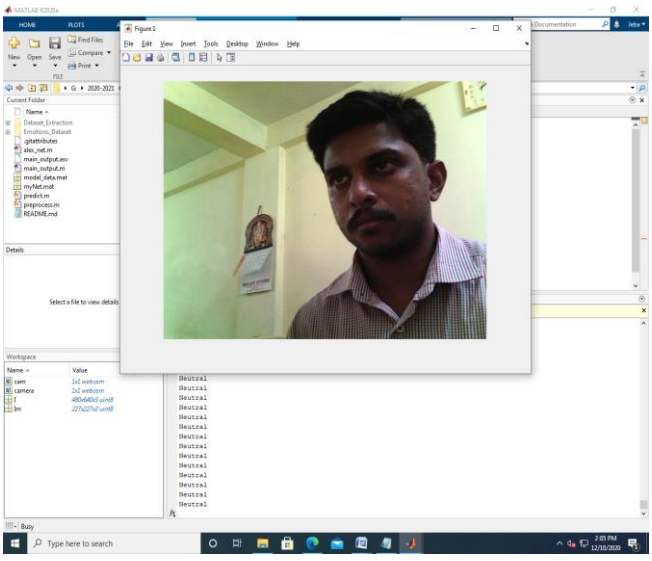


Resize Image

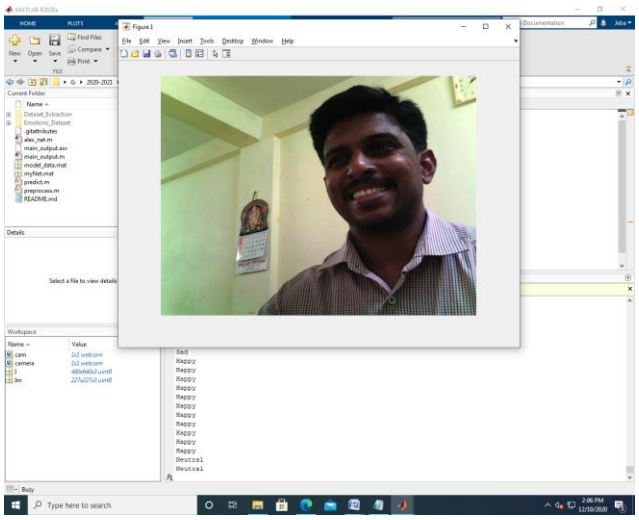


CNN Training

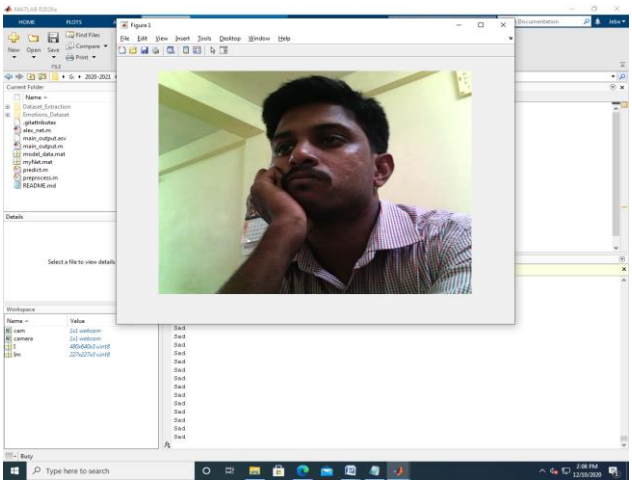
Original Image resolution



Face Expression Neutral



Face Expression Happy



Face Expression Sad

IV. CONCLUSION

In this work we present CNN with attention mechanism for facial expression recognition. The system recognizes faces from students' input images using Haar-like detector and classifies them into seven facial expressions: surprise, fear, disgust, sad, happy, angry and neutral. The proposed model achieved an accuracy rate of 70% on FER 2013 database. Our facial expression recognition system can help the teacher to recognize students' comprehension towards his presentation. In our future work we will focus on applying Convolutional Neural Network model on 3D students' face image in order to extract their emotions and we will study how to generate attention parts in faces without landmarks, as CNNs rely on robust face detection and facial landmark localization modules.

REFERENCES

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in CVPRW. IEEE, 2010, pp. 94–101.

[2] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in ICME. IEEE, 2005, pp. 5–pp.

[3] G. Zhao, X. Huang, M. Taini, and S. Z. Li, "Facial expression recognition from near-infrared videos," Image & Vision Computing, vol. 29, no. 9, pp. 607–619, 2011.

[4] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality- preserving learning for expression recognition in the wild," in CVPR. IEEE, 2017, pp. 2584–2593

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," arXiv preprint arXiv:1708.03985, 2017

[6] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated cnn for occlusion aware facial expression recognition," in 2018 International Conference on Pattern Recognition (ICPR), 2018.

[7] . Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE TPAMI, vol. 31, no. 2, pp. 210– 227, 2009.

[8] E. Osherov and M. Lindenbaum, "Increasing cnn robustness to occlusions by reducing filter support," in CVPR, 2017, pp. 550–561.

[9] 9. 9. M. Ranzato, J. Susskind, V. Mnih, and G. E. Hinton, "On deep generative models with applications to recognition," in CVPR. IEEE, 2011, pp. 2857–2864.

[10] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," Image and Vision Computing, vol. 26, no. 7, pp. 1052–1067, 2008.

[11] R. G. Harper, A. N. Wiens, and J. D. Matarazzo, Nonverbal communication: the state of the art. New York: Wiley, 1978.

[12] 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. 124-129, 1971.

[13] C. Tang, P. Xu, Z. Luo, G. Zhao, and T. Zou, "Automatic Facial Expression Analysis of Students in Teaching Environments," in Biometric Recognition, vol. 9428, J. Yang, J. Yang, Z. Sun, S. Shan, W. Zheng, et J. Feng, Éd. Cham: Springer International Publishing, 2015, p. 439-447.

[14] A. Savva, V. Stylianou, K. Kyriacou, and F. Domenach, "Recognizing student facial expressions: A web

application," in 2018 IEEE Global Engineering Education Conference (EDUCON), Tenerife, 2018, p. 1459-1462.

[15] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions," IEEE Transactions on Affective Computing, vol. 5, no 1, p. 86-98, janv. 2014.