The Process of Facial Expression Classification from Videos

Danyal Quazi
School of Computing
Dublin City University
Dublin, Ireland
danyal.quazi2@mail.dcu.ie

Khizer Ahmed Biyabani
School of Computing
Dublin City University
Dublin, Ireland
khizer.biyabani2@mail.dcu.ie

***Abstract*—Identifying emotion using facial expressions is among the most exciting research fields. Much work is being done in classifying emotions from images, while videos seem to be ignored. This study is a series of experiments attempting to classify emotions from a collection of videos annotated with the help of a survey. There are broadly three approaches; the first employs a Convolutional Neural Network, and the second combines a Histogram of Oriented Gradients (HOG) feature descriptor and ML models like SVM, KNN and Random Forest. The third approach uses an off-the-shelf, open-source python library, 'FER', to classify the emotions from videos. FER-2013 and CK+ datasets are used to train models separately. Finally, the results from the experiments are compared to find the techniques that perform better on videos and images. It observed that models perform significantly better on images compared to videos. The paper discusses the challenges that arise when dealing with video data and possible reasons that cause accuracy to fall when videos are analyzed compared to images.**

***Index Terms*—Emotion Recognition, Convolutional Neural Network, Facial Expression Recognition, Computer Vision, Face Detection, Image processing, Data Augmentation, Pre-processing, Histogram of Oriented Gradients (HOG), Support Vector Machine(SVM), K-nearest neighbour(KNN), Random Forest(RF), Precision, Accuracy, Recall, F1-score, Data Annotation, Confusion Matrix**

I. Introduction

In today's society, emotions are crucial in our everyday lives since they help decision-making, learning, and communication with others. Facial expression recognition is one of computer vision's most extensively debated and exciting problems. It offers a variety of benefits and applications in fields including medical, notably in mental health systems, education, and various recommender systems.

Facial recognition and Body language are two of the most critical aspects of communication and are vital factors in whether the communication is effective or successful. It is easier for a human to express rather than explain using words. When presented with an image or a video, humans are exceptionally good at understanding or interpreting these expressions or body movements, while computers or machines see nothing but a collection of 0s and 1s. Converting this into something meaningful is a real challenge.

Ekman, in a study, postulated that a genuine representation of an individual's inner emotional condition is facial expressions. A set of six so-called fundamental emotions (fear, anger, surprise, disgust, sadness, and happiness) are universally expressed and genetically encoded in humans and animals. Reading eyes, the furrow of the forehead, and even the curving of our lips serve as a medium and source of information about the feelings they are expressing. [1]

There are several approaches to extracting information from an image, each with its benefits and drawbacks. As discussed further, this paper starts by going over the techniques used in facial expression identification utilizing neural networks, feature descriptors with ML algorithms, pre-processing methods to improve accuracy and many more discussed in the next section. It is observed that compared to the work done in images, little work and resources are available for videos.

The third section of the paper focuses on the data used in this research, providing details about the image datasets used for the training and testing the proposed techniques. Two datasets are used in this study: FER-2013 and CK+, which are discussed in detail in the dataset section. A survey was conducted to create a dataset to compare and evaluate the techniques on how they perform on videos. This was achieved by using some creative common videos and annotating the videos with seven fundamental emotions using human annotators. The dataset section also presents a detailed explanation of the process of creating this dataset.

The paper then explains each approach attempting to classify a video's emotion in detail. The first solution uses a Convolutional Neural Network, which uses the FER-2013 and CK+ datasets for training, generating two separate trained models. The trained models are then used to predict the expressions in the testing images. Finally, the trained CNN models predict the most dominant emotion inside the annotated videos from the survey.

Similarly, the second approach uses Histogram of Oriented Gradients (HOG) and Machine Learning algorithms like Support Vector Machine (SVM), K-nearest neighbours (KNN) and Random Forest Algorithm. The HOG converts the images into an array of length 900 which is used to train ML algorithms. The best-performing algorithm is used to predict the emotions of the videos.

Furthermore, we have another approach using python's library FER, which can also be employed to analyze our annotated videos. Using FER to classify the emotion, we can compare the performance of other approaches on videos. Finally, in the evaluation section, the performance of all the techniques on all three datasets is compared using various evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

Finally, we analyze and assess our findings and the future work that can be done to improve the performance. We also discuss various challenges when working on videos and why models perform better when working with images than videos.

II. Literature Review

A lot of work is done in the field of Facial Emotion Recognition from static images. There are several techniques used to understand emotions. We can broadly split the steps into four categories: pre-processing, Feature Extraction, selection of a model and evaluation.

A study proposed a method using images or video clips and data from the internet. The world wide web is a vast source of knowledge; one study [2], for example, uses photos and videos from a variety of sources. The basic approach is with help of a 2D image we are able to extract information of a 3D video frame [3]. Identically in the discipline of deep neural network, employing machine learning and artificial intelligence algorithms, sentiment analysis has resurfaced as one of the most prominent fields in recent years, and when talking about deep neural network, CNN comes into play. Another study uses The Facial Expression Recognition 2013 (FER-2013) dataset coupled with a deep learning model to recognise seven different emotions. [3]

Another study focuses more on the pre-processing stage of the entire process and proposes an effective facial expression recognition system. The study discusses various challenges, such as light intensity variations, as well as how to deal with them. The proposed system employs Histogram of Oriented Gradients (HOG) extractor and Support Vector Machine (SVM) classifier. The advantages of HOG over Local Binary Patterns are discussed in this work (LBP). On the JAFFE database, the proposed system achieved an accuracy of 97.62 percent, while on the Cohn-Kanade database, it obtained a precision of 98.61 percent [4]. There are certain limitations that also follow in this approach as stated, that facial muscle movements are not related to a specific identifiable emotional experience in a one-to-one manner. Instead, emotions are most likely created in the mind of the perceiver, and emotional mental categories are required to appropriately categorize face movements amid surrounding data. [5]

There was another interesting approach where an Expression classification system was proposed in which Multi-Task Convolution Neural Network (MTCNN) was implemented. The method used not only facial expressions but also upper body movement or gestures to classify human emotions in this study. The system is divided into two branches. The Facial Expression Recognition branch is the first, while the upper body movement or gestures branch is the second. These branches were ultimately combined to get better results and make the entire system more robust. This innovative system was able to achieve an accuracy of 99.79% which proved how effective it is [6]. This can be very helpful when understanding the theme of a movie as we can combine the facial emotions with the upper body movement.

Apart from this, a study [7] works on classifying only two emotions, happy and sad. But the actual aim of this system is to detect genuine emotions. This system can classify emotions as acted or genuine. The proposed system shows how even simple algorithms can give exceptionally good accuracy if a correct dataset is used. The face was detected using the Cascade Classifier method from the OpenCV library. To differentiate the acted and natural emotions, three algorithms were used, kmeans was implemented to analyze the database and make clusters. K-Nearest Neighbour and Backpropagation Neural Network was used to classify the emotions.

A study estimated the emotional state of surveillance targets without knowledge about the presence of a firearm. Affective states elicited by carrying a firearm might be reflected in changes in the individual's body language. It is therefore possible that, when attempting to detect the carrier of a concealed firearm, the observers would respond to the change in non¬verbal behaviour of the bearers by attributing different affective states to surveillance targets. [8]

In [14] more stress is given to real-time emotion classification. In this study, facial landmarks and electroencephalograph (EEC) signals were used. The classifiers used in this study were Convolutional Neural Network (CNN) and long short-term memory (LSTM). For facial expression, ten landmarks were placed on the face to note the small muscular movements. The distance of these landmarks from the center of the face can be used to note the movement.

The [9] focuses more on the cinematographic content explaining its complexity and the problems which can arise due to the heterogeneity in the datasets. It also provided few solutions to deal with these problems and conducted various experiments.

There is an approach where the model selects some representative frames from videos rather than working on every single frame. The aim is to increase the accuracy and reduce the processing time. This is achieved by using only the frames where the face is captured from the front which contains sufficient information for recognition. [13]

Dalal and Triggs first introduced the HOG description in 2005. It detected pedestrians with complicated backgrounds almost perfectly. Recent years have seen a rise in the employment of HOG descriptors coupled with SVM classifiers in numerous applications for reliable object detection and recognition. [15] [16]

Movies or videos are complex but contain much information and, if handled properly, can be very helpful. A lot of work is done in the field of Facial Emotion Recognition from static images. Many studies deliver impressive accuracy in detecting basic facial emotions from images. However, there is no reliable dataset for specific tasks that can be used when working with videos. This study aims to find a Facial Expression Recognition approach that can perform better on videos.

Not only that, the study attempts to find various challenges and limitations in the facial Expression classification task on videos. This study also aims to understand how human annotators perceive emotions in a video compared to how a trained ML model analyzes the facial expressions in it.

III. Dataset

A. The FER-2013 Dataset

We have used Facial Expression Dataset, the dataset contains labelled images resembling 7 emotions, as shown in the table below:

TABLE I
THE 7 EMOTIONS AND THEIR IMAGE COUNT IN FER-2013 DATASET

EMOTION	IMAGE COUNT
Happy	8989
Angry	4953
Disgust	547
Fear	5121
Neutral	5928
Sad	6077
Surprise	4002

With the same emotions mentioned above in table 1, the dataset is separated into a training and a test set. There are a total of 28709 training photos and 7178 testing and validation images. This is used to train the models, which is discussed in the further sections. [20]

B. The Extended Cohn-Kanade Dataset (CK+)

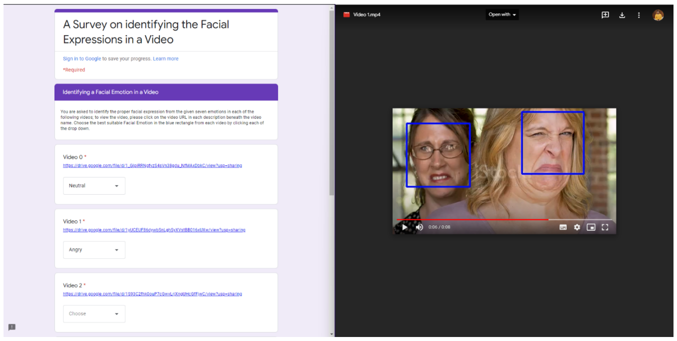
The Extended Cohn-Kanade Dataset (CK+) is another popular dataset for expressions. It comes with a total of 981 images which are annotated with seven basic emotions. Unlike FER-2013, the dataset is not already divided into training and validation sets. The dataset size is significantly smaller but much cleaner and has almost negligible annotation mistakes. The details of the data are provided in the table below. [18]

TABLE II
THE EXTENDED COHN-KANADE DATASET (CK+)

Emotion	Number of Videos
Happy	207
Angry	135
Disgust	177
Fear	75
Neutral	54
Sad	84
Surprise	249

C. Labelled Video Dataset

This research project aims to predict facial expressions from a video clip. While there was no easily accessible or publicly available labelled data for videos to achieve our desired results, hence we performed an expression elicitation task in which five different annotators labelled each of these videos with the seven basic emotions, namely, ’fear’, ’neutral’, ’sad’, ’happy’, ’anger’, ’surprise’ and ’disgust’. Each individual’s consent was taken along with a plain language statement, using a google form as shown in the figure below. The ethics form was also submitted and approved by the ethics panel at Dublin City University.



A Survey on identifying the Facial Expressions in a Video

Sign in to Google to save your progress. Learn more

*Required

Identifying a Facial Emotion in a Video

You are asked to identify the proper facial expression from the given seven emotions in each of the following videos; to view the video, please click on the video URL in each description beneath the video name. Choose the best suitable Facial Emotion in the blue rectangle from each video by clicking each of the drop down.

Video 0 *
https://drive.google.com/file/d/1_GlpiRRNofvzS4sVn38gdu_NfMAxDbkC/view?usp=sharing
Neutral

Video 1 *
https://drive.google.com/file/d/1yUCEUF86dywbSnLghSyXVstBB016xUXw/view?usp=sharing
Angry

Video 2 *
https://drive.google.com/file/d/1S93C2fhk0ouP7cGwvLriXnqUHcGfFjwC/view?usp=sharing
Choose

Fig. 1. Snip from Google Form used from the survey

The aim of this dataset creation was also, to compare the perspective of the human annotators towards these videos and the machine perspective in understanding the expressions from these videos. These creative commons videos were chosen at random from a public domain [19]. A total of 105 videos were collected, 40-50 seconds each; the table below shows the number of videos collected for each expression.

TABLE III
LABELLED VIDEO DATASET

Emotion	Number of Videos
Happy	26
Angry	20
Disgust	7
Fear	5
Neutral	23
Sad	16
Surprise	8

Five annotators labelled the expression for each video. The representative label for a video is decided by selecting the most given label from the five entries. It is calculated by taking the mode of the five labels given to a video. The exact process is repeated for all the videos generating the ground truth dataset; This dataset can be used later to validate models.

IV. Methodology

A. Emotion Classification using CNN

Convolutional neural networks are a kind of feed-forward neural network that are used in image analysis. It is employed to identify and categorize objects in images. ConvNet’s job is to reduce the images into a more straightforward format to analyze without eliminating elements essential for obtaining an accurate result. ConvNet outperforms other classification algorithms because it requires less pre-processing of the data.

The following figure shows the basic architecture of the convolutional neural network implemented in this research. There are three different layers: the input, the hidden, and the output layer; the hidden layer is where the convolution takes place.

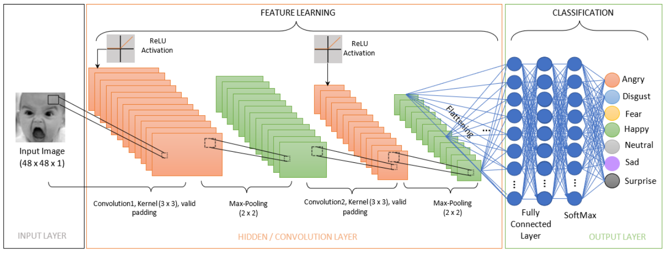


Fig. 2. CNN Architecture

The input is an image of 48x48 pixel. As mentioned earlier two data-sets were used for training two separate CNN models.These trained models are then used to predict the label from the videos. The process is described in the figure below:

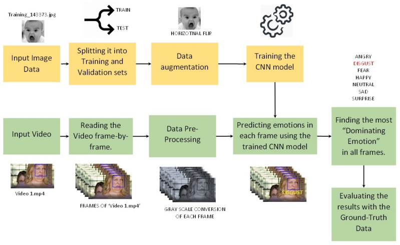

Training_143373.jpg
TRAIN
TEST
HORIZOTNAL FLIP
Input Image Data
Splitting it into Training and Validation sets
Data augmentation
Training the CNN model
ANGRY
DISGUST
FEAR
HAPPY
NEUTRAL
SAD
SURPRISE
Input Video
Reading the Video frame-by-frame.
Data Pre-Processing
Predicting emotions in each frame using the trained CNN model
Finding the most "Dominating Emotion" in all frames.
Video 1.mp4
FRAMES OF 'Video 1.mp4'
GRAY SCALE CONVERSION OF EACH FRAME
DISGUST
Evaluating the results with the Ground-Truth Data

Fig. 3. Process Flow Diagram

The input image data is initially split into training and validation sets. Now, to train the model, we are training them in batches of 64 photos throughout the cycle. The data is first processed using several data augmentation techniques, which prevents the neural network from learning irrelevant patterns, thereby improving its performance. Several data augmentation techniques, such as horizontal flip, which rotates the images by 180 degrees, re-scaling the images and many more are used for this purpose. Pythons' keras function ImageDataGenerator was used for this purpose.

Moving on to the hidden layer or convolution layer of the network. The Convolution Operation's goal is to take the input image's high-level characteristics, such as edges, and extract them. Typically, low-level features like edges, colour, gradient direction, etc. are captured by the first ConvLayer. With more layers, the architecture adjusts to High-Level characteristics as well, giving us a network that comprehends the dataset's images holistically, much like we do. This is done by multiplying the input weights by a two-dimensional array of weights referred to as Kernel/Filter.

Next to every convolution we have a pooling layer, it is resposible for dimensionality reduction, this reduces the computing power and aids the training process in allowing the extraction of dominating characteristics. In this case we have used the max-pooling method, that uses the input pools maximum value. Adding to that, to avoid overfitting we have added a dropout layer. There is rectified linear function (ReLU) that is used in the convolution process. It is used to rectify the no linear input image data into a linear one. There are a lot on non-linear features in an image such as- the transition between pixels, the borders, the colors, etc.

After going through the approach outlined above, we train our model with the input image features. Next, we will flatten the output for classification purposes and feed it into a standard neural network. The final layer contains the Dense Layer with the activation function also called as the Softmax. feed-forward neural network receives the flattened output, and back-propagation is used for each training iteration. The model can categorise images using the Softmax Classification method across a number of epochs by identifying dominant and specific low-level features. [12] The evaluation of the CNN model on each of the two datasets is discussed later on in this research.

Now that we have build a model that classifies expressions using an image, we now predict the emotions in the videos using trained CNN model.The videos can be imagined as a collection of frames or multiple 2D images, these frames are first converted into grayscale. Grayscale representations are frequently used for extracting descriptors instead of directly operating on colour images because they simplify the algorithm and reduce computational requirements.

These grayscale frames are now used to detect faces using OpenCV's Haarcascade Classifier, and the CNN model uses these detected faces to predict one of the seven emotions. The CNN model's performance is then evaluated using the labelled dataset's ground truth emotions, as discussed later in this text.

B. HOG and ML algorithms

In this approach, Histogram of Oriented Gradients (HOG) and Machine Learning Models were employed. The steps involved in the entire process are shown in the image below.

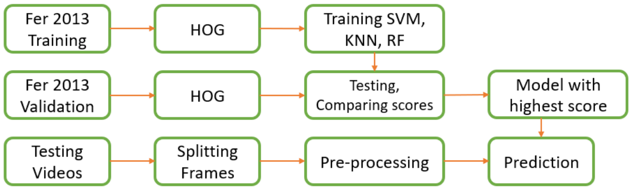


Fig. 4. Steps involved in the approach

1) HOG of Images: The images need to be represented using something simple like an array that will carry all the important information required for the further process and get rid of useless information. HOG converts an image with multi-dimension into an array of length n. This array can be used as the feature vector for the training of an ML algorithm. In HOG the distribution of directions of gradients is used as a feature. The HOG works on the fact that the corners of objects in an image carry a lot more information compared to the flat surfaces due to the lesser changes. [17] To calculate HOG, a python function hog is used from python's skimage library. In this approach images of size 48 X 48 are used which generates an array of length 900 when fed to the hog function. This function generates two outputs: actual hog features that can be used for the ML algorithms and a HOG image that can be used for the visualization. A list containing arrays each of length 900 was generated with another list containing labels. These two lists are further used for the training of the ML models. An example of a HOG image is shown in figure 5.

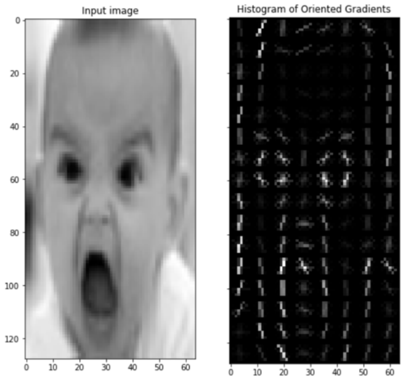


Fig. 5. HOG image example

2) Using ML Algorithms: The features extracted for all the images are then used to train Support Vector Machine (SVM), K-nearest neighbors (KNN), and Random Forest Algorithms. These models were trained on 28709 training images from FER-2013 belonging to 7 classes of emotion. The FER-2013 was already gray-scale images of size 48 X 48, hence weren't required to process. The two lists containing the features and labels respectively were given as the input for the training of the models.

Once the models were trained, all three models were tested using the validation set of the FER-2013 containing 7178 images belonging to the same class. Similar to the training set, HOGs of the testing set were calculated. The features were provided as an input to the trained model and were made to predict the labels. The results and scores are discussed in detail in the evaluation section. Compared to the other two, SVM performed better on the testing set and hence, was selected to predict the expressions in the video dataset.

The same process is repeated on the CK+ dataset. The CK+ did not have two separate datasets for testing and validation. The dataset was split into two smaller datasets containing 800 and 181 images for training and validation respectively. The results of this experiment are also discussed in the evaluation section. The SVM model trained on CK+ was also used to classify the testing videos.

3) Testing on Videos: Once the models are trained and tested on images, it was time to see their performance on annotated videos. The basic preprocessing tasks are common for CNN and this method, like splitting the video in frames. There are around 400 to 700 frames generated per video depending on the video length. Each frame is converted to grayscale using open CV's cvtColor() function. CascadeClassifier from Open CV is used to detect faces in the frame. The frame is then resized to 48 X 48 so that it matches the images on which the algorithm was trained. HOG for each frame is calculated and because the frames were resized to 48 X 48, it generates an array of length 900 same as the training data. For every video, the label for every frame is stored and the most repeated emotion is selected as the predicted emotion for the video. This task took more than 4 hours to complete for each trained model. The values obtained from these experiments are compared with the values given by the annotators to get the final results which are discussed in the evaluation section.

C. FER Python Library

The final method we have used is a python library called FER (Face Emotion Recognizer). It is a pre-trained library that can detect all the emotions present in an image file. It can classify all seven basic emotions and gives a score to each of them. The emotion with the highest score is considered the dominant emotion and represents the emotion in the image. When working with images, an image can be directly passed to the fer function which returns a dictionary containing the emotions with their scores. The model performs well when applied to images.

The model can recognize emotions from an image or a frame, to recognize the emotion of an entire clip, three steps are involved: Splitting the video file into frames, sending each frame in the video as an input to the classifier to get the emotion values of all the emotions in the frame and finally summing the emotion values of all the frames to get the final emotion scores of the entire video. The emotion with the highest combined score of all the frames is considered the dominant emotion of the video or simply the recognized emotion of the video.

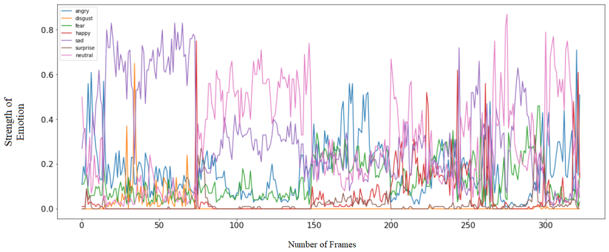


Fig. 6. Emotions throughout a video identified by FER.

The same steps are repeated for all the videos from the testing video dataset using a loop, only appending the dominant emotion of the video to a list. This list contains the emotions detected from testing videos and can be used to compare with the ground truth or the emotion tags given by the human annotators.

By default, the model uses OpenCV's Haar Cascade classifier for the detection of faces in a frame but in this study, MTCNN is used as it performs way better. This difference is due to the output bounding box of OpenCV's Haar Cascade classifier being square whereas the one with MTCNN is a rectangle that changes according to the face to cover the entire face making it more flexible. But this decision has its downside, using MTCNN for face detection makes the process slower. As mentioned in the paper, this research aims to find a solution for Emotion Recognition from videos. This makes the entire process a lot more time-consuming. To get results quicker and avoid the systems to stay on for hours, the testing videos were split into two separate datasets and were executed on two different systems. The results from these two datasets were later combined for evaluation. [11]

V. Evaluation and Results

Identifying facial emotions is a multi-class classification problem, as we have more than two possible classes (at present 7 emotion categories). There are several evaluation metrics that can be used for this purpose:

1) Accuracy: It simply counts the number of times the classifier correctly predicts the given facial emotion. The ratio of the number of correct emotions predicted to the total number of predictions made by the classifier can be defined as accuracy.
2) Confusion Matrix: It is a performance metric for classification problems with two or more classes as output. It is a table that contains a mix of predicted and actual values, as shown in figure below.

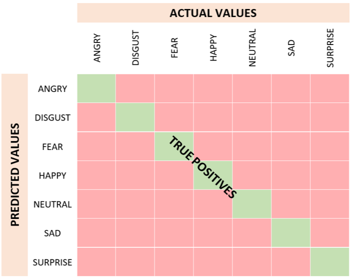


Fig. 7. Confusion Matrix

From a confusion matrix we further can determine:

a) Precision: Precision explains how many of the correctly predicted emotions actually turned out to be correct.
b) Recall: Recall explains how many of the actually correct emotions we were able to predict correctly with our model.
c) F1 Score: It is a combined result of both precision and recall.

$$\text{F1-Score} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

Each of the above mentioned metrics are used to evaluate the models, for classifying the facial emotions.

A. CNN Model

As discussed above we have trained the CNN model on two different image datasets FER-2013 and the CK+, and we evaluated their results. In the beginning we trained the model with 50 epochs through the neural network. The following figures 8(a) and (b) shows the training and validation loss and accuracy when trained on the two datasets respectively. It is evident from the graph that as the weights of the neural network are altered frequently the curve tends to shift from being underfitting to ideal to being overfitting,

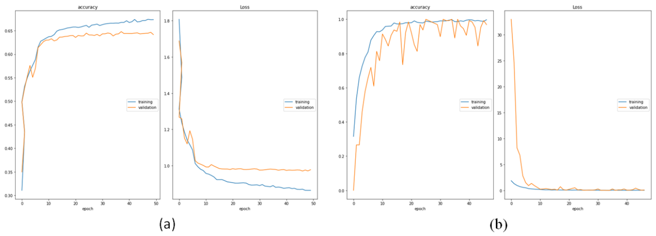


Fig. 8. Accuracy and Loss during training and validation of CNN using (a) FER-2013 (b) CK+ datasets

The CNN model which was trained on the Facial Expression Dataset (FER 2013) gave a validation accuracy of 0.64 and 0.96 when trained with the CK+ dataset, as shown in the graph. Now, we then evaluate the predictions of the model for the videos, the following table IV gives us the classification report for the trained CNN model on videos. The average accuracy for the CNN model trained on FER-2013 dataset was found to be 0.41 and when the model is trained with CK+ it gave an accuracy of 0.21. Also, figure 9 shows the confusion matrices for the same.

TABLE IV
Classification Report for CNN tested on Videos

	CNN (FER-2013)			CNN (CK+)		
	Precision	Recall	F1-score	Precision	Recall	F1-Score
Angry	0.5	0.25	0.33	1	0.1	0.18
Disgust	0.5	0.14	0.22	0.22	0.28	0.25
Fear	0	0	0	0.09	0.4	0.15
Happy	0.6	0.34	0.43	0	0	0
Neutral	0.38	0.73	0.5	0.21	0.47	0.29
Sad	0.32	0.62	0.42	0.18	0.125	0.14
Surprise	1	0.12	0.22	0.6	0.375	0.46
Accuracy	**0.41**			**0.21**		

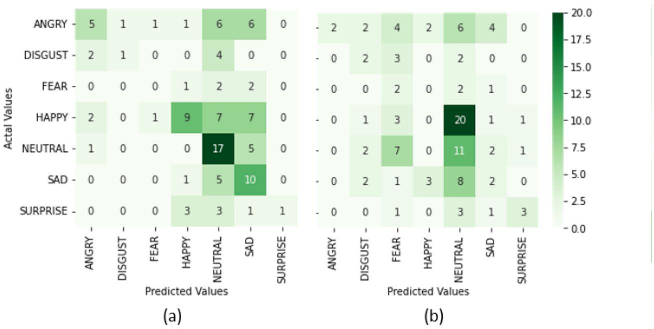


Fig. 9. Confusion Matrix for CNN tested on videos trained on (a) FER-2013 and (b)CK+

The CNN model trained on CK+ image data gave significantly better accuracy compared to when trained on FER-2013. On the contrary, the FER-2013 trained CNN model predicted better results for the videos as compared to the CK+ trained

CNN model. This anomaly in accuracy is discussed the next sections.

B. HOG and ML algorithms

Tables 5 and 6 show a detailed comparison of scores for evaluation matrices like recall, precision and F1-score for each emotion for FER-2013 and CK+ respectively. Average accuracy is also shown at the bottom of the tables for all three models.

It is clear from both the tables that models perform much better when trained and tested on the CK+ dataset compared to the FER-2013. The reason for this difference is the smaller size of the CK+ dataset and many wrongly labelled images in the FER-2013 dataset, which affects the testing and training of the models.

TABLE V
CLASSIFICATION REPORT FOR ML MODELS TRAINED ON FER-2013

	SVM			KNN			Random Forest		
	Precision	Recall	F1-score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Angry	0.36	0.43	0.39	0.40	0.32	0.36	0.21	0.41	0.47
Disgust	0.20	1.00	0.33	0.42	0.34	0.37	0.27	1.00	0.43
Fear	0.30	0.45	0.36	0.33	0.41	0.36	0.27	0.44	0.33
Happy	0.80	0.63	0.71	0.78	0.56	0.65	0.84	0.48	0.67
Neutral	0.54	0.48	0.51	0.42	0.42	0.42	0.45	0.40	0.43
Sad	0.45	0.41	0.43	0.22	0.45	0.29	0.29	0.36	0.32
Surprise	0.61	0.73	0.67	0.51	0.63	0.56	0.50	0.80	0.62
Accuracy	**0.53**			**0.47**			**0.47**		

It is also seen that SVM provides the best results for both datasets and hence is selected for prediction on the video dataset. The confusion matrices show the detailed classifications of all models performed for each label.

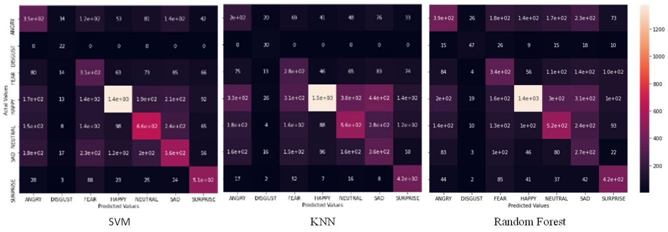


Fig. 10. Confusion Matrix for ML models trained on FER-2013 Dataset

When the models are tested on videos, it is evident that the accuracy falls significantly regardless of the dataset selected. The scores from the model trained on CK+ are almost useless, with only 0.21 accuracy, which is same for CNN model. Minimal examples for training in CK+ is the most significant factor as there are not enough examples for emotions in CK+ dataset to predict real-world scenarios.

TABLE VI
CLASSIFICATION REPORT FOR ML MODELS TRAINED ON CK+

	SVM			KNN			RF		
	Precision	Recall	F1-score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Angry	1.00	0.93	0.96	0.81	0.53	0.64	1.00	1.00	1.00
Disgust	1.00	1.00	1.00	0.69	0.6	0.64	1.00	0.97	0.98
Fear	1.00	1.00	1.00	0.84	0.57	0.68	0.84	1.00	0.91
Happy	1.00	1.00	1.00	0.65	0.82	0.73	1.00	0.93	0.96
Neutral	0.91	1.00	0.95	0.71	0.5	0.58	0.71	1.00	0.83
Sad	0.90	1.00	0.95	0.33	0.37	0.35	1.00	1.00	1.00
Surprise	1.00	0.97	0.99	0.62	1.00	0.76	1.00	1.00	1.00
Accuracy	**0.98**			**0.68**			**0.98**		

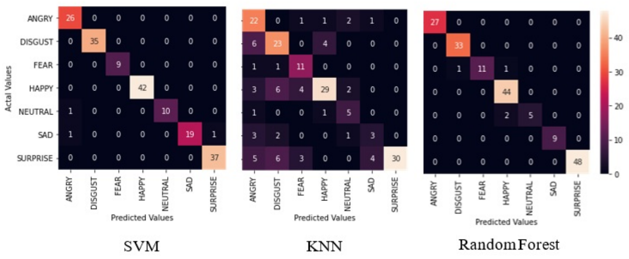


Fig. 11. Confusion Matrix for ML models trained on CK+ Dataset

Although SVM trained on FER-2013 scores less when tested on image dataset, it performs significantly better on video dataset thanks to the larger FER-2013 dataset, making the model more robust compared to CK+. The scores are still not very impressive, with an accuracy 0.40. The possible reasons for the lousy performance of models on videos are also discussed in the next section.

TABLE VII
CLASSIFICATION REPORT FOR ML MODELS TESTED ON VIDEOS

	SVM (FER 2013)			SVM (CK+)		
	Precision	Recall	F1-score	Precision	Recall	F1-Score
Angry	0.57	0.2	0.29	0.29	0.10	0.15
Disgust	0	0	0	0.06	0.14	0.08
Fear	0	0	0	0.27	0.60	0.37
Happy	0.85	0.46	0.6	0.82	0.35	0.49
Neutral	0.43	0.6	0.5	0.00	0.00	0.00
Sad	0.25	0.62	0.35	0.43	0.19	0.26
Surprise	0.5	0.25	0.33	0.08	0.59	0.13
Accuracy	**0.40**			**0.21**		

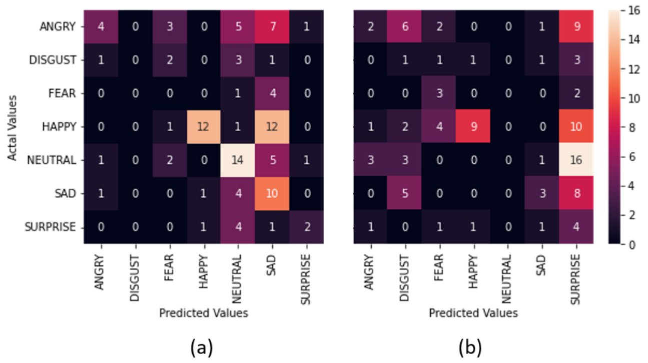


Fig. 12. Confusion Matrix for SVM on Videos trained on (a) FER-2013 and (b)CK+ datasets.

C. FER Python Library

Similarly, we used FER python API on each of our videos and evaluated the results, the following table 4 gives us the classification report in this case. The average accuracy for this model was found to be 0.57. The confusion matrix is as shown in figure 13.

The confusion matrix for both the models clearly explains the number of emotions correctly and incorrectly classified.

TABLE VIII
CLASSIFICATION REPORT FOR FER PYTHON API

Metrics	Precision	Recall	F1-Score
Angry	0.76	0.65	0.70
Disgust	0.00	0.00	0.00
Fear	0.00	0.00	0.00
Happy	0.78	0.81	0.79
Neutral	0.52	0.61	0.56
Sad	0.33	0.56	0.42
Surprise	0.75	0.38	0.50

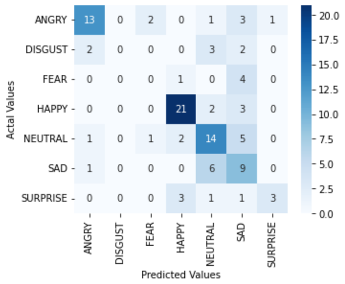


Fig. 13. Confusion Matrix for FER Python API

When comparing both the models we find that FER API was more accurate in predicting the basic emotions from the videos, as you can see for example for the emotion ’Angry’ the CNN predicted out of 20 only 5 of them correct, on the contrary the FER library was able to predict 13 of them correct hence giving better accuracy metrics.

VI. DISCUSSION

The paper has covered many steps involved in the process of emotion recognition from both images as well as videos, and it is clear from the results that the accuracy gained when working with videos is significantly less compared to the scores obtained in images; in fact, miles behind the better approaches mentioned in the literature review section. One of the significant reasons that this study was able to find which results in this fall in accuracy is the fluctuation in the facial expressions. The video is not just one static frame with one dominant emotion but hundreds of frames, each with entirely different emotion scores. The graph in figure 6 the FER section shows the fluctuation in emotions.

Different data augmentation techniques that take into account the facial posture or position of the face, can also be implemented. For instance, there are certain videos in our labelled video dataset where the emotions are not identified because the model was unable to detect any face in the video. These factors also hindered the performance in identifying the expression from the videos.

Another reason that causes the emotion recognition process to get complicated when working with videos is, the difference in how humans and machines perceive emotions for each facial expression. A machine works only on some fixed parameters, purely facial expressions. There are multiple standard parameters, for example, eye movement and lip movement. However, multiple factors are entirely different for both parties. A human understands emotion or an entire theme of the video. For example, there are videos in the testing data that contain multiple subjects. In this case, an algorithm works by splitting the video into frames that eliminate the motion (discussed later) and then separately classifies the facial expressions of each subject to add the scores to calculate the emotion mathematically. On the other hand, a human observes the interaction between the subjects, the action, reaction, etc. This causes the difference in the label given by a human and an algorithm.

During the study and the survey, it was noticed that motion is, in fact, a very crucial element when understanding emotions. The way a subject moves his/her head, the closing or opening of eyes, the contraction of facial muscles, the movement of the mouth, and many more. The first step when dealing with video is splitting it into frames and removing motion from the parameters. For example, there are videos in the testing set with a minimum facial expression where an algorithm will most likely label it as neutral, but a human would annotate it as angry as the theme is angry.

The point here is that we are dealing with facial expressions and the emotions they portray; we are not concerned with understanding the theme. These are the factors that significantly influence the scores when working with videos.

VII. CONCLUSIONS AND FUTURE WORK

This study successfully implemented three techniques to classify the expressions from videos: CNN, HOG+SVM and python’s FER API. FER-2013 and CK+ datasets were used for the training of these models. A survey was used to generate a dataset to evaluate the models successfully. The quality of the dataset used for training and testing is the most significant influencer that can affect the outcome. It was also observed that the size of the data is just as important as the quality of data, as proved in our experiments, that the models performed better on videos when they were trained on a larger dataset, that is, the FER-2013 dataset. Out of all three approaches, python’s FER API provides the best results.

It was also discovered that there were many mistakes in how the survey was conducted and the type of videos that were selected. As discussed in the previous section, there is a significant difference in how humans and machines deal with video data, and there several factors that make videos complicated to deal with.

To sum up, there is still a lot of room for improvement in this area of research, especially when working with videos.

ACKNOWLEDGMENT

The authors of the paper would like to thank Dr. Graham Healy for supervising the entire research and the annotators who contributed to annotating the videos. Finally, we would also like to thank Dublin City University for giving us an opportunity to work on this research.

References

[1] Tcherkass of, A., Dupré, D. The emotion–facial expression link: evidence from human and automatic expression recognition. Psychological Research 85, 2954–2969 (2021). https://doi.org/10.1007/s00426-020-01448-4

[2] Wang, S., Shen, X. Zhang, Y. 3D facial feature and expression computing from Internet image or video. Multimed Tools Appl 77, 22231–22246 (2018). https://doiorg.dcu.idm.oclc.org/10.1007/s11042-018- 5895-7

[3] S. K. Lalitha, J. Aishwarya, N. Shivakumar, T. Srilekha and G. C. R. Kartheek, ”A Deep Learning Model for Face Expression Detection,” 2021 International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), 2021, pp. 647-650, doi: 10.1109/RTEICT52294.2021.9573626.

[4] S. Roy Supta, M. Rifath Sahriar, M. G. Rashed, D. Das and R. Yasmin, ”An Effective Facial Expression Recognition System,” 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECONECE), 2020, pp. 66-69, doi: 10.1109/WIECONECE52138.2020.9397965.

[5] Ekman, P. (2017). Facial expressions. In J.-M. Fernández-Dols J. A. Russell (Eds.), The science of facial expression (pp. 39–56). New York: Oxford University Press.

[6] Zaghbani, S., Bouhlel, M.S. Multi-task CNN for multi-cue affects recognition using upper-body gestures and facial expressions. Int. j. inf. tecnol. (2021). https://doi.org/10.1007/s41870-021-00820-w

[7] J. Rannow Budke and M. Da Costa-Abreu, ”Using neural and distance-based machine learning techniques in order to identify genuine and acted emotions from facial expressions,” 11th International Conference of Pattern Recognition Systems (ICPRS 2021), 2021, pp. 121-126, doi: 10.1049/icp.2021.1433

[8] 1. Blechko A, Darker IT, Gale AG. The Role of Emotion Recognition from Non-Verbal Behaviour in Detection of Concealed Firearm Carrying. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2009;53(18):1363-1367. doi:10.1177/154193120905301842

[9] Almeida, J.; Vilaça, L.; Teixeira, I.N.; Viana, P. Emotion Identification in Movies through Facial Expression Recognition. Appl. Sci. 2021, 11, 6827. https://doi.org/ 10.3390/app11156827

[10] Facial expression detection using Machine Learning in Python, Published in Analytics Vidhya by Aaditya Singhal Jan 5, 2021. url: https://medium.com/analytics-vidhya/facial-expression-detection-using-machine-learning-in-python-c6a188ac765f

[11] The Ultimate Guide to Emotion Recognition from Facial Expressions using Python, Published in Towards Data Science, by Rahulraj Singh Jul 26, 2021 url: https://towardsdatascience.com/the-ultimate-guide-to-emotion-recognition-from-facial-expressions-using-python-64e58d4324ff

[12] A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, Published in Towards Data Science, Dec 15, 2018 by Sumit Saha url: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[13] I. S. Topkaya and N. G. Bayazit, ”Improving face recognition from videos with preprocessed representative faces,” 2008 23rd International Symposium on Computer and Information Sciences, 2008, pp. 1-4, doi: 10.1109/ISCIS.2008.4717905. url: https://ieeexplore.ieee.org/document/4717905

[14] Aya Hassouneh, A.M. Mutawa, M. Murugappan, Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods, Informatics in Medicine Unlocked, Volume 20, 2020, 100372, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2020.100372.

[15] N. Dalal and B. Triggs, ”Histograms of oriented gradients for human detection,” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[16] T. Surasak, I. Takahiro, C. -h. Cheng, C. -e. Wang and P. -y. Sheng, ”Histogram of oriented gradients for human detection in video,” 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 172-176, doi: 10.1109/ICBIR.2018.8391187.

[17] Histogram of Oriented Gradients explained using OpenCV, Satya Mallick DECEMBER 6, 2016, LearnOpenCVhttps://learnopencv.com/histogram-of-oriented-gradients/

[18] 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,” 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[19] Source for Videos: https://www.pexels.com/license/

[20] The source of FER-2013:https://www.kaggle.com/datasets/msambare/fer2013

IEEE