CMP304: Artificial Intelligence Unit 2 Report
Emotion Recognition

Calum Gillies
1903400

Introduction and Background

Facial recognition has become an increasingly important area of development in various industries including healthcare, entertainment, and marketing. Identifying the human emotion of the detected face also has its benefits. It can be extremely useful for adding an extra layer of security where both the face and the emotion of the subject is extracted. To 'verify that the person standing in front of the camera is not just a 2-dimensonal representation,' is another use case (Gilligan et al, no date).

Facial recognition in the marketing industry is an effective use of AI technology. If companies had the opportunity to capture and detect the real-time emotions of a consumer based on image or video capture, then conglomerates would be able to make an informed decision of whether a product was received well (Gilligan et al, no date).

Stated in 'Facial Emotion Recognition Using Machine Learning,' Raut states that human emotions can be classified as: 'fear, contempt, disgust, anger, surprise, sad, happy and neutral' (Gilligan et al, no date).

Identifying objects or faces in images falls into the category of image classification. This means the AI can be trained to classify and assign a label to an image it receives input for (Sheridan, 2022). In regard to categorising images to specific groups through feature extraction, researchers and developers use this technique for a multitude of tasks, not only for faces. In the medical field, image classification is already being utilised for identifying cancerous regions of the skin by having a neural model compare cancerous vs non-cancerous images, although the technology is still in its infancy (Goyal (et al)).

Other researchers focus on developing models that can perform the same tasks for facial structures using neural architectures such as Convolutional Neural Networks (CNN) in a way to understand what parameters or training methodologies affect the confidence, learning rate and accuracy. In the paper 'Emotion AI, Real-Time Emotion Detection using CNN,' the researchers focus on these aspects while exploring different models and how they perform on datasets put through different processing techniques including grayscale conversion and gaussian blurs (Gilligan et al, no date).

This project focuses on emotion recognition through the extraction of facial structures in images. Gathering large datasets of real-world imagery depicting people of a wide range of genders, races and in varied environments is a costly endeavour, so for the purpose of this analysis, a dataset of cartoon characters has been used instead. These images follow a similar standard of emotion labelling as above. For this project, a Support Vector Machine will be used.

Implementing facial recognition is a time-consuming process where hyperparameters and a varied testing dataset must be tweaked and prepared in order to see a pre-trained model perform well on new data that it may not be strictly designed for.

In this scenario, the model will be trained on cartoon expressions and so it is potentially not going to be nearly as effective at classifying emotions.

The aim of this project is to supervise the learning of a model using the OpenCV library to identify the emotions present in images of animated characters. The efficacy of this will be tested by occluding some of the image features to see how this affects performance, while also drawing comparisons between the accuracy of utilising feature engineering versus using a list of landmarks detected.

Data Specification

The dataset used in the project was obtained from the Facial Expression Group Database (FERG-DB) (Aneja et al, 2016). This is a collection of roughly 15,000 images pertaining to two characters, each displaying the following emotions: anger, disgust, fear, joy, neutral, sadness and surprise.

The image dataset was divided into training, validation, and testing segments. While training the agent, the training data would be used while supervising the learning process. Validation is the stage where new unseen images are introduced to confirm that the agent is 'learning' the underlying patterns of the images instead of merely memorising the images from the training set. Generally speaking, this is the stage the hyperparameters will require fine tuning to prevent overfitting and underfitting. The testing segment will be used as an overspill of images that can be used to increase the validation or training image count if this can show an improvement in the model.

The dataset is structured by dividing an animated male and female character into separate folders, each containing 7 folders labelled with the corresponding emotions. The images inside the folders range from slight to extreme expression. This means that the data sample is highly varied and so results taken can be used to represent the entire data set as a depiction of the overall performance. This should be true excluding the training data since the model will have reference of the exact landmark locations on the detected faces.

Data Generation

Using Google Colab, a dataset of landmarks was generated by presenting the model with the training data segment. Through supervising the learning process here, the model is given access to all of the training images where it is told specifically how to classify the image. The emotion label is extracted from the corresponding folder and appended to the emotion column in the dataset file as are the positions of all 68 features extracted. The x and y components of the features are included in this file, giving a total of 136 features values.

	A	B	C	D	E	F	G
1	emotion	feature(1)	feature(2)	feature(3)	feature(4)	feature(5)	feature(6)
2	anger	69	153	69	167	71	181
3	anger	69	153	69	167	71	181
4	anger	65	158	67	172	70	185
5	anger	65	158	67	172	70	185
6	anger	65	158	67	172	70	185
7	anger	70	148	70	163	72	178
8	anger	68	149	69	165	71	180

Figure 1

A second landmark dataset file was also generated by making correlations between specific facial features that the model extracted. This reduced the amount of datapoints from 136 to 6 in total which represents the facial features such as eyes, lips, nose, and mouth. In this case the features focused on were as follows: left eyebrow, right eyebrow, left lip, right lip, lip width and lip height.

	A	B	C	D	E	F	G
1	emotion	LeftEyebrow	RightEyebrow	LeftLip	RightLip	LipWidth	LipHeight
2	anger	5.991441327	2.933127556	5.111700919	5.226354224	2.995094028	0.69026849
3	anger	6.181662817	2.90317924	5.690474756	5.976406535	3.441549403	1.27072975
4	anger	5.959401634	2.872397577	5.452069922	5.440169415	3.243029661	0.913500278
5	anger	6.279351308	2.767596338	5.069076331	5.356451959	2.841913192	0.766964989
6	anger	6.088212713	2.851269263	4.892190985	5.087263515	2.708325198	0.64122345
7	anger	6.670316191	2.774640595	5.396876487	5.585918396	3.239838225	0.913500278
8	anger	6.001885839	2.842898095	6.08865152	6.284604019	3.712142239	1.403566885

Figure 2

This dataset was generated by providing the model with the same training data for detecting facial structures in the same way as before.

Since an animated dataset is being used, the accuracy of the values seen in the figures above may not be completely accurate but can still provide insights.

Figure 3 - Technique (Munasinghe, 2018)

The red circles in *Figure 3* represent the features extracted using the feature engineering approach. This image in particular is useful for seeing the accuracy of the model when not identifying real-world facial structures.

Methodology

The program was focused on extracting all of the features of a face using a publicly available model. This pre-trained shape predictor was loaded into the project so it could be used for processing the datasets. Initial testing was conducted on this image to see if the code cell was set up correctly and was in fact extracting the features.

Figure 4 - (Kanade et al, 2000)

A list is containing 68 parts (x and y components of each landmark feature) is when the extraction is successful. Since using animated faces may cause inaccuracies, a real-life image was used instead. Upon successful implementation of the face detection, the next step was to read through the training data and capture the features.

The learning is supervised here by parsing through the training set of images. Each image has the face detector applied to it and the resulting features have their x and y values unpacked before writing to a .csv document. We give the emotion label to the model so it will have a reference of the emotion in regard to the values detected and written to the dataset.

SVM Algorithm

This project will use the Support Vector Machine algorithm. It is a binary classification algorithm which aims to find an appropriate decision boundary (hyperplane) in a high-dimensional feature space that separates different classes of data points i.e. 'anger' and 'joy.' It is extremely effective when parsing linearly separable datasets. By using a different model function, the SVM can use non-linearly separable datasets but for this project, linear data is being used. The ability to handle high-dimensional feature spaces and its effectiveness in binary classification tasks is the reason why it is the chosen algorithm. The SVM will be trained

on the extracted facial features and the corresponding emotion labels to learn the underlying patterns and make predictions on unseen data.

Classification using SVM (Support Vector Machine)

Using the generated data based on the training images, a SVM is then used to classify the validation set of unseen images. The model parameters were then set accordingly. The kernel was set to use the 'linear' functionality because the data is stored linearly as can be seen in *Figure 1* and *Figure 2* above.

The value of 'C' when setting the model up relates to the significance of each data point to focus on when extracting information. If this value is increased then the model can begin to overfit the data which, in this case, essentially means the model is memorising the data rather than the underlying patterns of each image, leading to exaggerated accuracy outputs. The 'C' value can be lowered but this can cause the opposite problem where the model cannot learn effectively and will potentially produce inaccurate performancs. This is why the use of the validation set is crucial because more reliable results can be taken from the unseen images.

After setting up a model and running it on the data, it was tested using the initial generated dataset file against the validation set. A visualisation was then produced as a scatter plot.

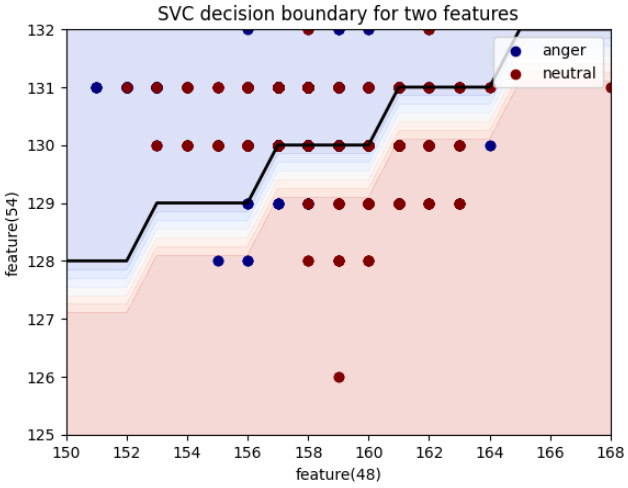

Figure 5

This figure shows the relationship between the emotions of 'neutral' and 'anger'. Compared features here are located on each side of the lips (54 and 48). These features are actually used to calculate the width of the lips and so play a big part in the identification of emotions.

Feature Engineering

In the hopes of making more meaningful predictions from the extracted information, the program was modified to make use of collating data from selected landmark features. Here is the list of features and how they were calculated. (Note: The landmarks start at 0 and not 1, so all landmark features shown in this figure should be considered as 1 above the true value).

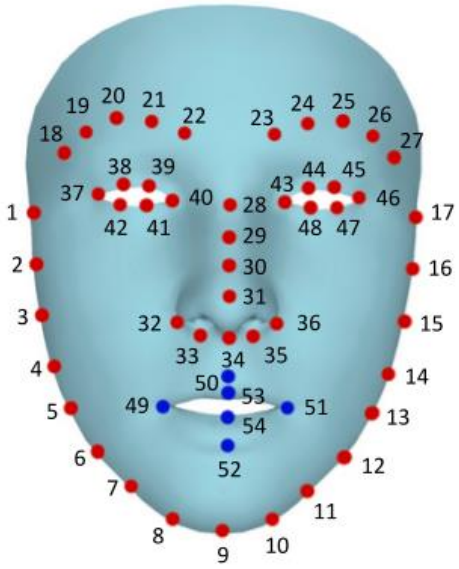

Figure 6 - (Kuzdeuov, 2021)

- Left Eyebrow: This was calculated by finding the distance between points [18], [19], [20], [21] and the inner eye point [39]. These four distances are each divided by the distance between [39] and [21]. This normalizes the values in relation to the size of the detected face. The total of the distances is then added to the landmark features list to give one complete feature.
- Right Eyebrow: Following the same structure as before, the new points used were [22], [23], [24], [25] and the inner eye point is [42].
- Left Lip: The new points were [48], [49], [50], [51], the stationary point being [33].
- Right Lip: The new points were [52], [53]. [54], [51], the stationary point still being [33].
- Lip Width: The width requires the distance between [48] and [54]. It is normalised by dividing this value by the distance between [34] and [51].
- Lip Height: The height is found by dividing the distance between points [51] and [57] by the distance between points [33] and [51].

When finding the distances, a method was implemented to recolour the circles on a test image (see fig above) for each feature, so they could be verified as the correct to use. Once this was working for the test image, a new dataset file was generated by parsing through the training images again and extracting the more meaningful landmarks, appending the relevant emotion label to it. This was then tested against the validation images in the same way as before.

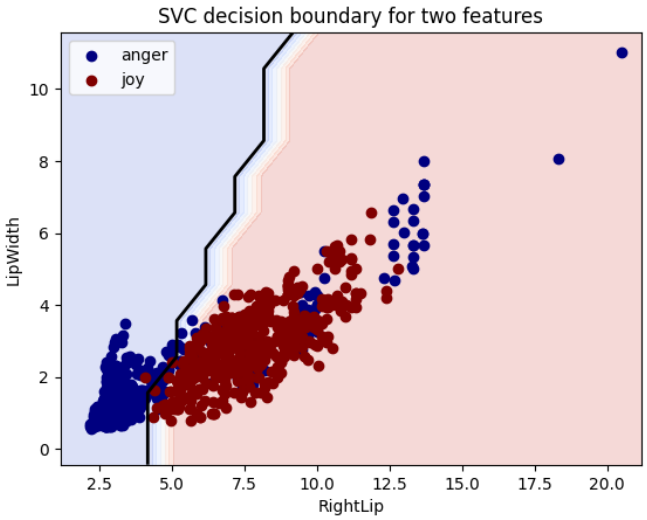

Figure 7 - Decision Boundary - Feature Engineering

Scaling the input data (data sets)

```
#Scaling the input data
sc_X = StandardScaler()
x_train = sc_X.fit_transform(x_train)
x_test = sc_X.fit_transform(x_test)
```

The scaling here is being applied separately to the testing and training set when setting up the model hyperparameters. This can lead to some inconsistencies and therefore present inaccuracies regarding the model's performance. When applying scaling to the two sets, it should be done consistently. Using the StandardScaler, scaling must be performed on the training set, but the same scaler needs to be used on the testing data also.

This is the amended code:

```
#Scaling the input data
scaler = StandardScaler()

#Fit the scaler on the training data
scaler.fit(x_train)

#Scale the training and testing data
x_train_scaled = scaler.transform(x_train)
x_test_scaled = scaler.transform(x_test)
```

Results and Conclusions

Note: Across all testing, there are times where a frontal face image is not detected as containing a face. For accuracy's sake, the testing will count these uninterpretable images as being a false prediction. The kernel will be set to the 'linear' model for all testing. Using 'poly'

or ‘rbf’ (radial basis functions) do not fit the data well. This analysis uses linearly separable data so ‘linear’ it is the suitable choice.

Test 1 – Landmark Features (68 Part List)

Conditions: The data is new unseen images from the validation set which has a total of 3790 images showcasing 7 emotions across two characters. In this test the dataset that contains a reference to each of the 68 returned landmarks is being used. A ‘C’ value of 1 was used.

Results:

Total Predictions:	3790
Correct Predictions:	3547
Images with no found features:	127
Accuracy:	93.59%

Figure 8

Discussion:

This is a high accuracy which is expected from a pre-trained shape detector. There were 127 images which the detector could not locate a face and so have been considered to be incorrectly predicted. Considering the high accuracy, it is a possibility that the model is overfitting the data, but this cannot be the case. The images used here are new and the model is not being instructed to retain the information observed in the images. Lowering the ‘C’ value of the model does not significantly impact the classification performance, suggesting that the decision boundary between the classes is not highly sensitive to the exact positioning of individual data points. Therefore, it can be stated that the model is relatively robust to small changes in the training samples and can still achieve a good accuracy by maintaining a wider margin. This also shows the efficacy of using the ‘linear’ property since the data is designed to be linearly separable.

Test 2 – Feature Engineered (6 Features)

Conditions: The features extracted are as follows: left eyebrow, right eyebrow, left lip, right lip, lip width, lip height. The validation set containing 3790 images are used; All 6 engineered facial landmarks are being used for the extraction of particular areas; A ‘C’ value of 1 is being used for prediction.

Results:

Total Predictions:	3790
Correct Predictions:	2191

Images with no found features:	234
Accuracy:	57.81%

Figure 9

The code for this test differs where each facial image is not only being passed to the face detector. The image is then passed to separate functions that are tasked with identifying specific landmark features using parts from the landmark list.

Discussion:

The execution time is much higher than the previous test, sitting at around 10 minutes which is double the time the first test took. This can be related to the extra processing required to identify specific landmark features on the face before extracting the calculated engineered features.

By using only 19 out of the 68 individual landmark parts for creating features, the dramatic drop in accuracy can be accounted for. The accuracy dropped to 57.81%, from the 93.59% in the previous test. Observing this result suggests that the engineered features used in this test may not be capturing the relevant information or patterns necessary for accurate emotion detection. Using the animated faces can be viewed as culpable because the features are exaggerated, be it the eyes, nose, or facial shape and so this discrepancy can make it more troublesome for a pre-trained facial recognition model to identify facial structures here.

Further Optimisation:

It is difficult to source a database of real-world images of people from various ethnic backgrounds which is suitable for emotion detection. The Cohn-Kanade (Kanade et al, 2000) images are a good starting point for developing a model to recognise emotions, but the set is not large enough to give meaningful results.

If further use of the animated images is to be used then it is strongly recommended that there be further features engineered for extraction, such as the nose and eyes as well as the facial contour.

Comparing Features

Comparing both techniques of feature extraction can give some interesting insights to the emotion recognition process by depicting the identification process:

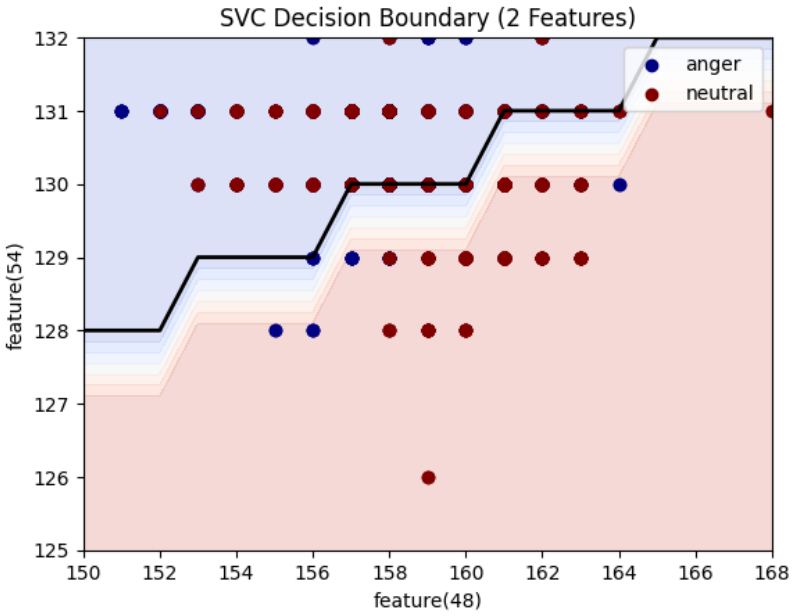

Figure 10

Landmark Features (Method 1) Analysis:

Looking at the lips in relation to the 'anger' and 'neutral' classes can provide some insights to how the Support Vector Machine decides on what emotion it predicts to be displayed in an image, and how effective it is at identifying an emotion using the mouth as a feature.

The scatter plot above shows feature 54 on the y-axis and feature 48 on the x-axis. These points are used for engineering the lip width feature, and so by taking these points from the list of landmarks used in the initial method, the relationship between these features and the emotions detected can be focused on.

The colour coding of the points, with blue and red representing 'anger' and 'neutral' respectively are displaying a cluster forming uniform rows, with each row containing a different number of points. The boundary line cuts almost evenly between the classes. Looking at the points on each side, there is more neutral on the 'anger' side, versus a lot less 'anger' plot points present on the neutral side. This suggests that the SVM is better at distinguishing neutral emotions compared to anger emotions. It also provides some evidence that when using a specific feature, the mouth can cause some misidentification regarding emotions.

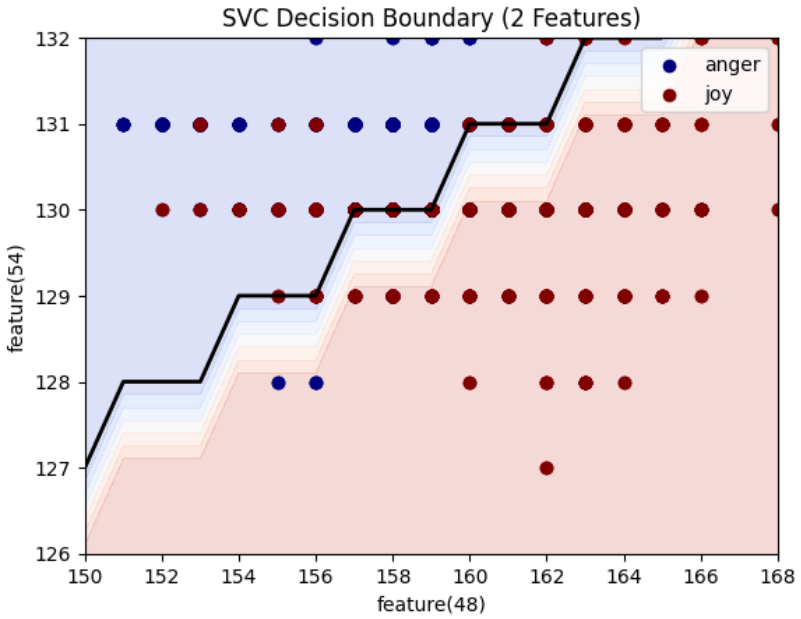

Figure 11

For example, looking at this graph shows a strikingly similar distribution of points regarding the lips. The same features have been used here but 'neutral' is swapped for 'joy'. If we have a look at the images with detected emotions, it can be proposed that the slightest change of a muscle in the mouth can completely change the emotion. This would decrease the efficacy of the SVM model and explain why there is slight overlap with the emotion labels plotted here.

Test 3: Occlusion Testing

Based on the findings above, testing was carried out on removing some of the features of the detected face. This can help to gain insight to the weighting of specific features, and how its removal or integration can affect the accuracy of the SVM.

Left Eyebrow Removal:

Total Predictions:	3790
Correct Predictions:	2021
Images with no found features:	234
Accuracy:	53.32%

Figure 12

Both Eyebrows Removed:

Total Predictions:	3790
Correct Predictions:	1687
Images with no found features:	234
Accuracy:	44.51%

Figure 13

Both Eyebrows & Lip Width and Height Removed:

Total Predictions:	3790
Correct Predictions:	527
Images with no found features:	234
Accuracy:	13.91%

Figure 14

The expected result occurred here where the consistent removal of features causes the model's efficacy to dramatically drop. In the validation image set there is a consistency of the number of images where a face cannot be detected. This suggests it may be the same images causing problems during each session, and in which case, they do affect the accuracy. In this instance, 234 images account for 6.17% which is significant. In future sessions these should be removed, or different images should be used instead. Due to the nature of using animated images, this is an issue to keep in mind throughout testing.

The decreasing of accuracy when removing the left eyebrow drops from 57.81% to 53.32% which is only slight. When both eyebrows have been removed it drops to 44.51%. A further dramatic decrease in the specificity of emotions drops to 13.91%, rendering the model totally unreliable.

The most interesting detail that can be deduced from these results is the removal of the lips width and height features. These features require normalisation to consider the scale of the features across different faces in images. When these features are removed, the SVM model struggles to accurately identify the emotions depicted. This suggests that the lip width and height features play a crucial role in emotion detection and should be given significant weight in the model.

Additional Features Implemented:

Observing the tables, the accuracy of the model still seems to be low but as said previously, the engineered features could be performing worse due to the shape predictor not being fed a real-world image.

To see if the efficacy of the model improves by using more of the total 68 landmark features, two more functions were added to check for the eyes in each image.

Total Predictions:	3790
Correct Predictions:	2287
Images with no found features:	234
Accuracy:	60.34%

Figure 15

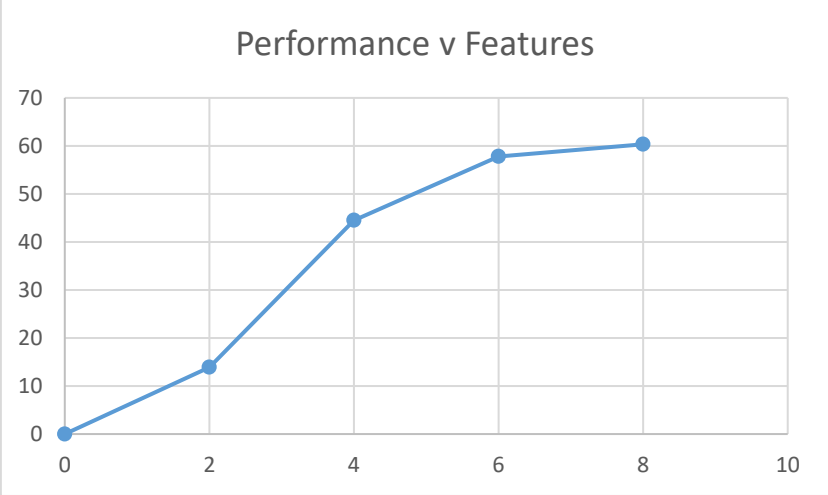

Figure 16

Overall, the feature engineering is a fruitful one where the efficacy of the model can be attributed to how much detail is included in the landmarks list. Adding the functionality to extract the eye features of the face increased the success rate by 2.53%. By adjusting the hyperparameters such as the 'C' value, the datapoint significance can be increased. By setting the value to 20, the model accuracy increases by 1.34%. Changing it to 30, there is another increase of 1.02%.

It is promising that the more features added, the higher the accuracy. This makes the model scalable. The nose, ears and facial contours were not coded for during the testing and may help to significantly boost the performance. Additionally, the altering of the data point importance does not cause the model to become incompetent when giving new images to process.

Since there are a finite number of landmark features that can be detected, it makes sense to include some further pre-processing steps such a implementing a grayscale conversion (Gilligan et al, no date). Conversion can make all images more uniform and also emphasise some features of the faces, making it easier for the shape predictor to extract specific features. It simplifies the images, reducing the complexity of the input data. A gaussian filter can be applied which may also aid in the extraction process. This can help smooth out noise and enhance the important details in an image (Gilligan et al, no date).

The raw data of the images can be passed to the face detector model, so it has more information to learn patterns from.

There are CNNs (Convolutional Neural Network) which are specifically tooled towards image-based tasks. They can automatically learn relevant features from raw image data, potentially capturing more intricate patterns and improving the accuracy of emotion detection. Comparing the performance of the SVM model with feature engineering and a CNN model would certainly provide more valuable information.

References

Gilligan, T., Akis, B./Stanford University(no date)Emotion AI, Real-Time Emotion Detection using CNN. Available at: https://web.stanford.edu/class/cs231a/prev_projects_2016/emotion-ai-real.pdf (Accessed: 28th April 2023)

Sheridan, S. (2022) Image classification in AI: How it works, Levity. Available at: https://levity.ai/blog/image-classification-in-ai-how-it-works#:~:text=Image%20classification%20is%20the%20task,multi%2Dlabel (Accessed: 02 May 2023). In-text: (Sheridan, 2022)

Goyal, M., Knackstedt, T., Yan, S., Hassanpour, S./Elsevier(2020)Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. Available at: https://reader.elsevier.com/reader/sd/pii/S0010482520303966?token=75222CEA15CBD49CA21F86F7E2271E8757B1B937B469B9FAE7FAC12490DAFC7911244B6CAC92E4F9F1BFEA80C4304B35&originRegion=eu-west-1&originCreation=20230521101515(Accessed: 4th May 2023)

Aneja, D., Colburn, A., Faigin, G., Shapiro, L. and Mones, B., (2016). Modeling stylized character expressions via deep learning. In Asian conference on computer vision (pp. 136-153). Springer, Cham

Munasinghe, M.I.N.P./University of Moratuwa(2018)Facial Expression Recognition Using Facial Landmarks and Random Forest Classifier. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466510(Accessed: 10th May 2023)

Cohn-Kanade AU-Coded Facial Expression Database: Kanade, T., Cohn, J.F. and Tian, Y., 2000, March.

Kuzdeuov, A. (2021) IS2AI/thermal-facial-landmarks-detection: SF-TL54: Thermal facial landmark dataset with visual pairs., GitHub. Available at: https://github.com/IS2AI/thermal-facial-landmarks-detection (Accessed: 19 May 2023).