Human Computer Interaction using Facial Recognition: An Exploratory Analysis of Aff-Wild2 Dataset

Tushara Govinda Reddy
Queen Mary University of London
Mile End Rd, Bethnal Green, London E1 4NS
t.govindareddy@se21.qmul.ac.uk

1. Introduction

Human Emotions have always been a vital part of non-verbal communication, that encouraged Automatic Facial Expression Recognition (FER) field of study to become the epicentre of many elaborate research projects given its vast applications in understanding human behaviour, identifying mental disorders, security, lie detection and synthetic human expressions [1]. Facial Expressions as such contribute to 55

2. Related Work

Traditionally, Dynamic Facial Expression Recognition analysis by Maja et al, classify human emotions as prototypic expressions of a person which as happiness, sadness, surprise, shock, anger and disgust based on their frontal view. Here the identification is purely done using frontal facial images, running the risk of not being able to handle temporal dynamics of facial actions.[2][4] Maja et all attempt to handle larger range of human facial expressions by recognizing the facial muscle actions that product different expressions. They propose a system for automatic recognition of facial Action units(AUs) and their temporal models from long, profile-view face image sequences. They detect the 15 facial points in an input face-profile sequence to introduce facial-action-dynamics recognition from continuous input video. The utilization of an algorithm to perform automatic segmentation of input videos into facial expression images (6 expressions) and recognise their temporal segments of 27 Action Units[2] occurring independently or combined with input video. Maja et al's method results in a recognition rate of 87

As per Jyoti et al [1], Automatic Facial Expression Recognition is accomplished using the following 5 steps represented in a diagrammatic representation as below in Fig1. Firstly, the pre-processing step converts the time-series of images from neutral to an expression, the facial component identification detects region of interest (ROI) for the action units (AU) such as cheeks, eyes, nose, eyebrows, lips etc. The feature extraction step deals with the extraction of these ROI's. The mostly popular feature extraction techniques are Local Binary Patterns, Local Gradient Code, Local Directional Pattern, Histogram of Gradient Orientations, Principal Component Analysis, Linear Discriminant Analysis. Support Vector Machine and Nearest Neighbour are the classification methods used for facial expression classification. LGC best improves the Recognition Rate (87.5

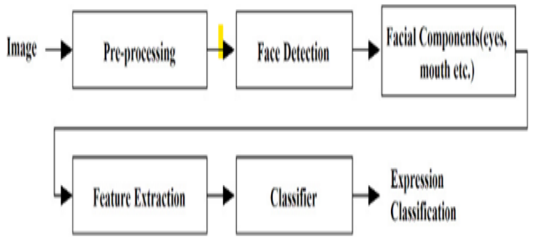


Fig.1. Facial Expressions Classification

Now we can dwell on some deep learning architectures built in FER on the Aff-Wild and similar image datasets. Kollias et al built CNN architectures together with RNN's for the Aff-Wild2 Challenge. Nested CNN models were trained on various datasets with facial and random images.[7] RNN and CNN are jointly trained in a specific sequence to obtain significant results on RECOLA[6] and Aff-Wild2 datasets[8]-[12]. This method of training is expensive with consumption of high computational resources and with a high complexity. These issues are overcome by using the attention mechanism which deals with the problem of the limited shot-term memory of RNN's.

Denis et all proposed a model where CNN model acted as a feature extractor and RNN was used to train the model

for facial recognition by exploiting temporal dynamics of the video unit to predict the valence and arousal score. The feature extractor is based on CNN architecture[12] with few modifications. To achieve a trade-off between underfitting and overfitting, we chose this architecture due to its simplicity compared to other complex onesTo extract relevant feature vectors from images, CNN is trained separately from RNN. RNNs can then learn useful information about temporal dynamics from sequences of feature vectors. RNNs can then learn useful information about temporal dynamics from sequences of feature vectors. Instead of using large frames, it allows large batches with small feature vectors, resulting in better performance.

Continuous affect prediction is a challenging and interesting problem in the wild due to the heavy computation involved. Sowmya Rasipuram et al[13] present the methodologies and techniques used to predict continuous emotion dimensions i.e., valence and arousal, in ABAW competitions based on Aff-Wild2 data. The Aff-Wild2 database consists of videos in the wild labelled at frame level for valence and arousal. Multi-modal features (multi-modal) are extracted using state-of-the-art methods in our proposed methodology. The audio-video features are used to train a sequence-to-sequence model using Gated Recurrent Units (GRUs). Validation data with a simple architecture shows promising results. Overall valence and arousal of the proposed approach are 0.22 and 0.34, respectively, which are better than the competition baseline of 0.14 and 0.24.[13]

3. Hypotheses-Restrictions

Hypothesis1: The purpose of this research is to investigate if the imbalanced data ratio can be balanced using down sampling, stratification or other balancing techniques.

Hypothesis2: To build and train a CNN on Aff-Wild2 with appropriate pre-processing techniques to avoid over fitting.

Hypothesis3: To use various data augmentation techniques such as normalization, shearing, perturbation, variation in brightness, contrast, blur and rotate to make the training data as resistant to mis-classification as possible.

Hypothesis4: To further prepare the data for building a CNN-RNN model that supports the temporal dynamics of changing emotions in large videos.

Hypothesis5: To achieve a reasonably high recognition rate close to 87.5

4. Data Preparation

Data Description: Some of the problems in the field of automatic FER are tackled in the Affective Behaviour Analysis in-the-wild Competition, third edition, held with IEEE International conference on Computer Vision and Pattern Recognition, 2022. The competition was held as four sub-competitions all working on Aff-Wild2 dataset, which is a large scale database which is annotated in terms of valence and arousal, expressions and action units[4]. The database was generated through augmentation 558 videos of 458 subjects with around 2,800,000 frames, showing both subtle and extreme human behaviours in real-world settings. Four expert annotated the database in terms of valence and arousal, three senior experts annotated 63 videos with 398835 frames in terms of AUs 1,2,4,6,1,2,15,20,25. Another 7 were assigned to annotate 539 videos consisting 2,595,572 frames in terms of 7 basic expressions. All annotations have been performed in a frame-by-frame basis. The four challenges included i) uni-task Valence-Arousal Estimation ii) uni-task Expression Classification iii) uni-task action unit detection iv) Multitask Learning. [4] Aff-Wild2 database is the first of its kind annotated data for all main behaviour tasks(valence-arousal estimation, action unit detection ,expression classification), whereas the most others contain annotations only for one task.[5]

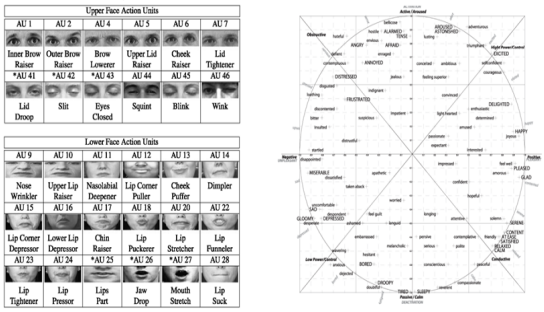


Fig. 1: The 2D Valence-Arousal Space & Some facial Action Units

Fig.4. Valence Arousal Estimation Classification and Action Units

Data Preparation: To create a machine learning lifecycle, the first most primary step of modelling is to split the data into training and training dataset.

1. Move to Dataframe: This files are are renamed with their class name concatenated with the file number and moved

	Automaker	Model	Images
0	ergasiaMLDqmul	ANGER	Anger3998.jpg
1	ergasiaMLDqmul	ANGER	Anger4003.jpg
2	ergasiaMLDqmul	ANGER	Anger4045.jpg
3	ergasiaMLDqmul	ANGER	Anger4035.jpg
4	ergasiaMLDqmul	ANGER	Anger4019.jpg
...	...	...	...
3995	ergasiaMLDqmul	SURPRISE	Surprise6561.jpg
3996	ergasiaMLDqmul	SURPRISE	Surprise6591.jpg
3997	ergasiaMLDqmul	SURPRISE	Surprise6566.jpg
3998	ergasiaMLDqmul	SURPRISE	Surprise6585.jpg
3999	ergasiaMLDqmul	SURPRISE	Surprise6577.jpg

21000 rows x 3 columns

Fig.3. Common DataFrame Created with renamed files

2. Duplicates: Raw data contains no duplicate annotated samples.

```
False    21000
dtype: int64
```

```
Empty DataFrame
Columns: [Automaker, Model, Images]
Index: []
```

Fig.4. No Duplicate labelling found

3.Training and Test Data split Splitting Data at 70:30 ratio into Train and Test respectively.

```
14700
6300
(14700, 2)
(6300, 2)
```

Fig.5. Splitting the Data into 70:30 ratio

4.Reshape: Reshaping is an essential step to carry out scaling of the pixel array.

```
(96, 96)
[[[[178]
   [182]
   [191]
   ...
   [199]
   [202]
   [204]]

  [[172]
   [163]
   [174]
   ...
   [205]
   [200]
   [203]]

  [[185]
   [179]
   [180]
   ...
   [199]
   [194]
   [215]]

  ...
```

Fig.6. Reshaping the Pixel Array

5. Balancing the data: It can be observed in the below histogram that Aff-Wild2 is a significantly unbalanced with the imbalance ratio of the majority class to the minority being 7:1, where the 'SADNESS' class has around 13040 samples and 'DISGUST' being lowest at 1858. The other classes fall between the two extremes('HAPPINESS'-9113,'SURPRISE'-7418,'ANGER'-6701,'FEAR'-2985).

Fig.7. Histogram Distribution of Sample Dataset

```
SADNESS      13040
HAPPINESS     9113
SURPRISE      7418
ANGER         6701
FEAR          2985
DISGUST       1858
```

1. Undersampling the majoirty class ”SADNESS” by removing samples to match upto a 6000-7000 mid-range.

2. Oversampling the minority class ” DISGUST”, ”FEAR” by adding more samples to match upto 6000-7000 mid-range.

3. Stratification: We stratify the data while dividing into train and test data to assume that equal ratio of sample population is picked in both. Thus reducing the risk of overfitting the model for training dataset of majority class.

4.Downsampling: Remove the frames with both values of valence and arousal equal to zero, then divide the values of bins into 40 bins for all samples in the dataset. Then a frame is selected for training with probability.

$$p = \frac{k_{current}}{k_{mf}},$$

where kcurrent is a number of samples in bin of this frame and kmf is a number of samples in bin of the most frequent frame

5. Data Pre-processing

Experimentation with Data Augmentation by Label Preserving and Perturbation carried out in the below sections. We employee colour augmentation to deal with noise, perturbation to deal with misclassification and normalization for scaling.

1. RBG to Grayscale: The main reason why grayscale representations are often used for extracting descriptors instead of operating on colour images directly is that GrayScale simplifies the algorithm and reduces computational requirements.

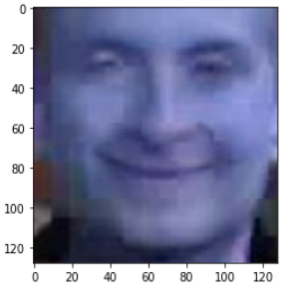


Fig.8a. Original Image in RBG Scale(128,128,3)

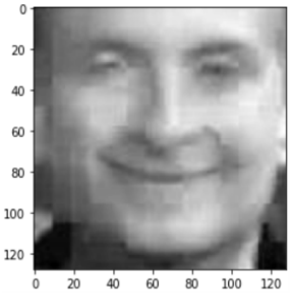
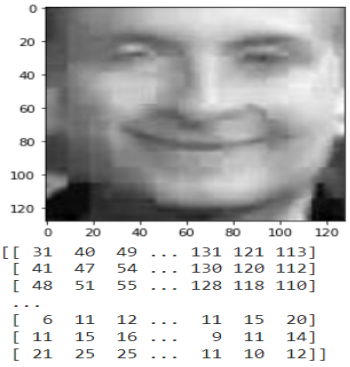


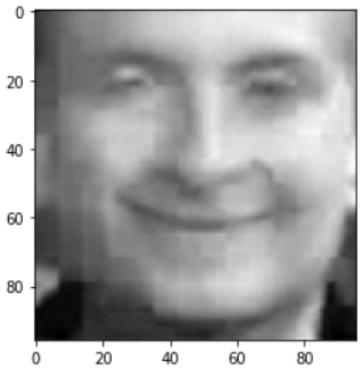
Fig.8b. Image down-scaled to GrayScale(128,128,1)

2. Normalization: We can observe that the RBG dataset ranges from 0 to 255 in form of (128,128,3) which indicates that in order of this data to be passed as inputs to the neural networks, they need to be of small weight values, because inputs with large integer values can disrupt or slow down the learning process. It is a good practice to normalise pixel values to have a value between 0 and 1. Before



```
[[ 31  40  49 ... 131 121 113]
 [ 41  47  54 ... 130 120 112]
 [ 48  51  55 ... 128 118 110]
 ...
 [  6  11  12 ...  11  15  20]
 [ 11  15  16 ...   9  11  14]
 [ 21  25  25 ...  11  10  12]]
```

Fig.9a. Original Image in GrayScale before Normalization

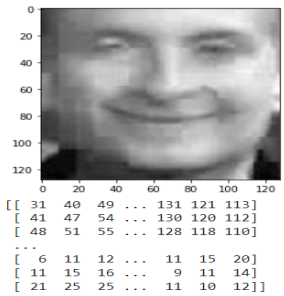


```
[[0.13333333 0.17647059 0.20392157 ... 0.53725349 0.49411765 0.44705882]
 [0.17647059 0.20392157 0.21960784 ... 0.5254902  0.48627451 0.43921569]
 [0.19215686 0.19607843 0.21176471 ... 0.51372549 0.47843137 0.43137255]

 ...

 [0.04313725 0.05490196 0.04313725 ... 0.0627451  0.06666667 0.08627451]
 [0.03529412 0.05490196 0.04313725 ... 0.04313725 0.04705882 0.0627451 ]
 [0.07843137 0.09019608 0.0745098  ... 0.05098039 0.03921569 0.04705882]]
```

Fig.9b. Image in GrayScale after Normalization

3. Resizing: Resizing enables better, quicker training of model without memory limitations prohibiting the training of CNNs at High resolutions.

Before

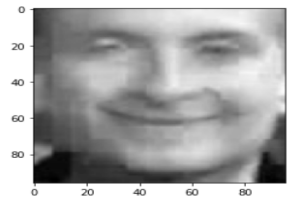


```
[[ 31  40  49 ... 131 121 113]
 [ 41  47  54 ... 130 120 112]
 [ 48  51  55 ... 128 118 110]
 ...
 [  6  11  12 ...  11  15  20]
 [ 11  15  16 ...   9  11  14]
 [ 21  25  25 ...  11  10  12]]
```

(128, 128)

Fig.10a. Original Image in GrayScale (128,128,1)

After



```
[[ 34  45  52 ... 137 126 114]
 [ 45  52  56 ... 134 124 112]
 [ 49  50  54 ... 131 122 110]
 ...
 [ 11  14  11 ...  16  17  22]
 [  9  14  11 ...  11  12  16]
 [ 20  23  19 ...  13  10  12]]
```

(96, 96)

Fig.10b. Resized Image in GrayScale (96,96,1)

4. Perturbation Introducing various types of noise to discover the minimum noise that can lead to model resulting in mis-classification.

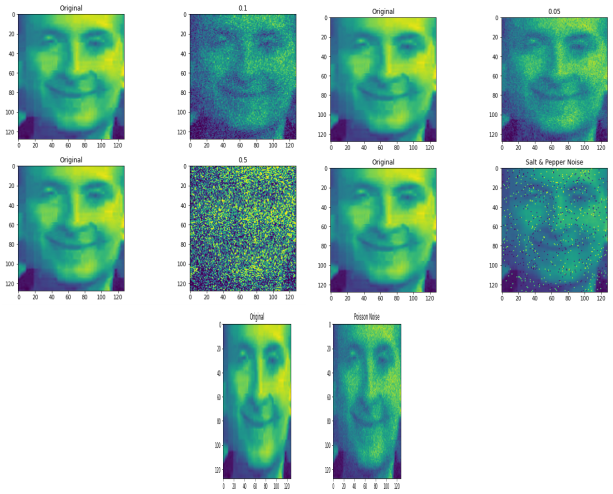


Fig.11. Perturbation by adding various noises

5. Shearing The resulting image undergoes shearing to provide a variety of data.

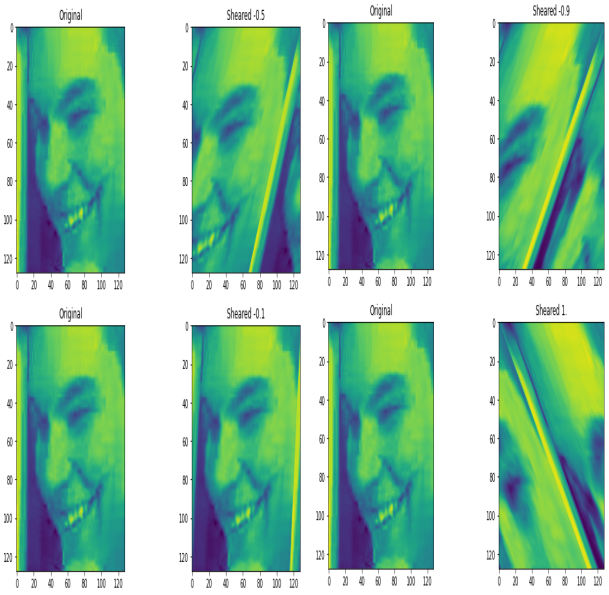


Fig.12. Data Augmentation by adding different shearing values

6. Blurring To remove noise and smooth-en out the edges we employee blurring technique.

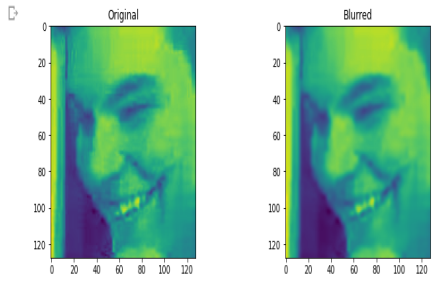


Fig.13. Label Preservation by blurring

7. Brightness The resultant image becomes darker or brighter relative to the original.

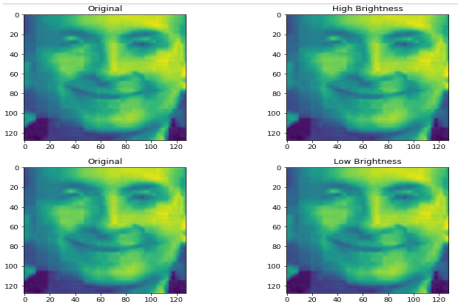


Fig.14. Colour Augmentation by fluctuating the brightness

8. Contrast The higher the contrast, the greater the difference between the background and the features to be inspected.

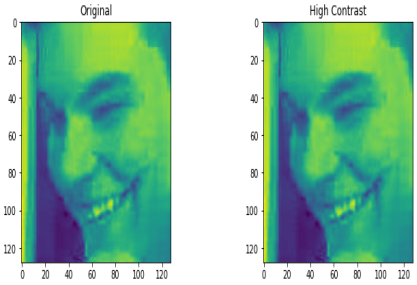


Fig.15. Colour Augmentation by fluctuating the contrast

9. Rotation Another label preserving technique for conversion from one coordinate space onto another.

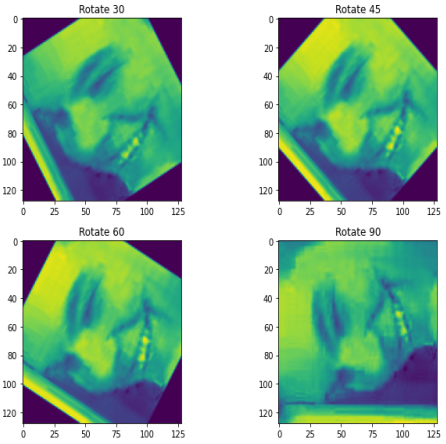


Fig.16. Label Preservation by rotation

References

[1] Kumari, J., Rajesh, R. and Pooja, K.M., 2015. Facial expression recognition: A survey. Procedia computer science, 58, pp.486-491.

[2] Pantic, M. and Patras, I., 2006. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36(2), pp.433-449.

[3]https://www.microfocus.com/documentation/idol /IDOL/Servers/IDOLServer/11.3/Guides/html/ English/expert/Content/IDOLExpert/ Improve/FaceRecognize$_factors.htm$

[4] Kollias, D., 2022. Abaw: Valence-arousal estimation, expression recognition, action unit detection multi-task learning challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2328-2336).

[5] Kollias, D. and Zafeiriou, S., 2021. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792.

[6] Rangulov, D. and Fahim, M., 2020, December. Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network. In 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS) (pp. 14-20). IEEE.7 D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," arXiv preprint arXiv:1811.07770, 2018.

[8] Kollias, D. and Zafeiriou, S., 2019. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855.

[9] Kollias, D. and Zafeiriou, S., 2018. A multi-task learning generation framework: Valence-arousal, action units primary expressions. arXiv preprint arXiv:1811.07771.

[10] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-thewild: Aff-wild database and challenge, deep architectures, and beyond," International Journal of Computer Vision, pp. 1–23, 2019.

[11] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017, pp. 1980–1987.

[12] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in Computer Vision and Pattern Recognition Workshop [13] Rasipuram, S., Bhat, J.H. and Maitra, A., 2020, November. Multi-modal Sequence-to-sequence Model for Continuous Affect Prediction in the Wild Using Deep 3D Features. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 611-614). IEEE.