Facial Emotion Recognition(FER): A Comprehensive Analysis and Survey of Emotion Recognition Techniques

Viswanatha Sharma C
Queen Mary University of London
Mile End, London, UK
v.s.chepuri@se21.qmul.ac.uk

1. Introduction

[4]The recent exponential rise in the deep learning technology with hardware, architecture has driven the facial emotion recognition into a new frontier of classifying in videos(wild images) and better static sequences of images that are being primarily labelled on the six basic emotions identified by Ekman et al which are anger, disgust, fear, happiness, sadness, and surprise. The use facial expression or emotion detection has prominent applications in the human computer interaction, medical treatment, fatigue and drowsiness surveillance of individuals while driving, hazardous workplace platforms etc.,

This paper provides a brief comprehensive review of various FER methods and models that were proposed and implemented in the past. In first part which is the discussion of related work, the paper goes through the various FER related methodology and critically analyze their results and describe the limitations of their dataset, architecture and model. Later, in the next section the paper delves into the real-world scenario application of the problem specified in those papers. This section will briefly explain these restrictions over the survey.

The data preparation section will go through the loading of the custom dataset a fraction of images from the Affwild2 dataset and Generated images obtained from said facial expression manipulation-transfer setting, and its properties with refining the dataset by deleting the duplicates, etc., discussing the results and statistics in before and after cases, and in the data Pre-processing will cover the image augmentations applied to images such as normalization and other techniques.

2. Related Work

The six basic emotions identified by Ekman et al which are anger, disgust, fear, happiness, sadness, and surprise [3]. There are other emotions added later such as neutral, embarrassment, contempt etc., This emotion detection is implemented with the sign- based algorithms to undergo training for detection of action units AUs in the given images. The use facial expression or emotion detection has prominent applications in the human computer interaction, medical treatment, fatigue and drowsiness surveillance of individuals while driving, hazardous workplace platforms etc.,

2.1. A neural-AdaBoost based facial expression recognition system [19]

This paper has performed face detection by implementing the Viola-Jones descriptor that is highly successful in facial detection with the use of Ada-boost algorithm that gives more weight to items that are classified incorrectly resulting in these having more significance next consecutive model. This process will not cease unless a low magnitude of error has been obtained i.e., all have same weights. The dataset used in the training and testing of model are JAFFE consisting of 213 images of different facial expressions from ten non related women of Japanese ethnicity and the other dataset being the YALE which consists of 165 grayscale GIF images of 15 different individuals. There are 11 images per subject [19].

While performing the image Pre-processing by applying Bessel down-sampling where the image size has been rescaled to pixels of 20 x 20 by the use of the images (crops) to the only required AUs such as forehead, eyes, nose, cheeks, mouth.The feature extraction method used is the Gabor Method that uses band pass filters in image processing for feature extraction, texture analysis. The Feature selection being inhand with respect to the gabor wavelets that have selected with the Ada-boost algorithm to speed the classification. The wavelets hypothesis that have highest discrimination for least classification error are picked and weights calculated. Later, obtaining the final feature selection hypothesis.

The classifier implemented is a Multilayer feed forward neural network (MFFNN) with architecture of 3- layer feed forward neural net and a trained back propagation algorithm for seven and six emotions to JAFFE and YALE respectively. The process of training involves weight initializa-

tion, calculation of the activation unit, adjustment, weight adaptation, and testing for convergence of the network. The accuracy of neutral recognition is poor and fear higher inferring that the muscle deformations around mouth and eyes on face is proportional to the automatic facial expression detection. The proposed classifier having better classification than other models. he average recognition rate in JAFFE database is 96.83% and that in Yale is 92.22%. The mild expressions have lower recognition rate. There is a non significant improvement of recognition rate for the YALE dataset of 2% to the SVM classifier. Implying that this method has not been able classify as expected for all real-world datasets. And, many further improvements are needed.

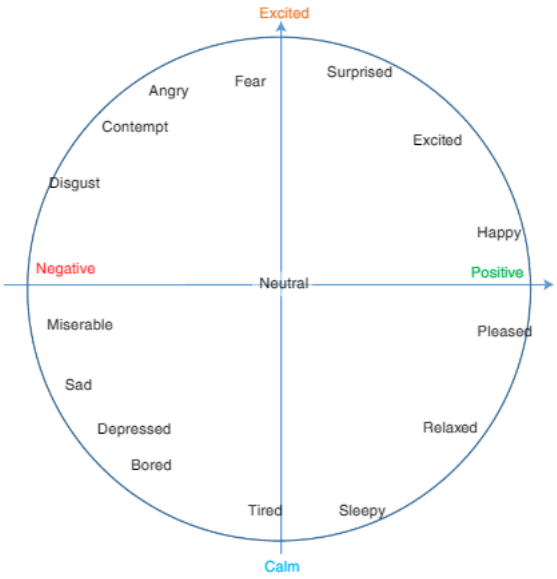
2.2. Facial expression recognition based on facial components detection and hog feature [2]

The he system detects the face first and then, extracts the facial components from the face image. After that, Histogram Oriented Gradient (HOG) is extracted to encode these facial components and concatenate them into a single feature vector. These feature vectors are used to train a linear SVM. considered the facial components and employed the HOG feature descriptors on the facial components attention to the facial components which contribute to the facial expression recognition. The one of the dataset used JAFFE. and other, The Extended Cohn-Kanade Dataset is used in the paper which consists of 123 subjects with 593 number of sequences, seven expressions being "angry, surprise,happy, disgust, fear, contempt, and, sad" and neutral expression. The Viola-Jones descriptor implemented for the determining the face region and its components. This features are encoded by HOG Histogram of Oriented Gradients to depict the appearance and properties of object with the distribution of the local gradient intensity and orientation. The obtained classification rates of this method 94.3% and 88.7±2.3%, respectively.

This method has higher classification for more visible expressions as anger, fear. However, relatively poor accuracies for mild expressions as contempt. That can be linked to the use of minimal feature extraction and simple SVM classifier. This can be optimized by implementing better preprocessing, The use of HOG has better feature extraction than the Gabor Transform. This accuracy may further increase with use of CNN based one like in [19].

2.3. Estimation of continuous valence and arousal levels from faces in naturalistic conditions [1]

This paper deals with the Real time motion images detection of the valence and arousal levels to specify it into one of the categories of Emotions



The single network with architecture named EmoFAN detects the facial key structural landmarks to emotions categorical and continuous estimation, implemented a attention mechanism to specifically focus on the regions for better classification. Dataset used are AffectNet, SEWA having satisfactory result exceeding other models performance. To handle the less size of the AFEW-VA dataset, the trained model EmoFAN from AffectNet has been fine-tuned for this with use of five-fold person-independent cross-validation strategy. The drawback of this method is to employ a dataset for training that consists of all variables in the naturalistic environment such ethnicity, gender, facial responses, behaviourial patterns, labeling. Because a discrimination is highly receptive and dependent on training data.

2.4. An SVM-AdaBoost facial expression recognition system [20]

The paper proposes techniques similar to [19]. However, the classifier that is used is an SVM support Vector Machine based classifier that considers the action units. Instead of Viola-jones seen in [19] an another MRC maximal rejection classifier that perform repetitive rejection operation over image to distinguish target and clutter, The DCT discrete cosine transform to transform inherent spatial domain existing image to the frequency domain, obtaining an ideal illumination on the face. Then the same Bessel downsampling and feature extraction by Gabor features given to the SVM. Trained on the JAFFE and YALE datasets. It has satisfactory results compared to other with recognition rate are 97.57 % and 92.33 % respectively. Showing better accuracy than that of [19].

2.5. Facial Expression Recognition Using 2DPCA on Segmented Images [5]

This paper proposes novel facial emotion recognition technique such that an image is segmented into parts with respect facial features such as left and right eye, nose, cheeks etc., then a 2-dimensional PCA principal component analysis based transformation. These obtained feature vectors are used for minimum distance classifier MDC that depicts the classification based on the euclidean distance. The classification accuracy for JAFFE dataset with proposed method is 89.86% and obtaining higher accuracy for mild Neutral emotion due to the method of segmentation.

2.6. Going deeper in facial expression recognition using deep neural networks [18]

The paper discusses a novel architecture implementing deep sparse networks, i,e., application of inception layer with two inception style modules made of 1×1, 3×3 and 5×5 convolution layers (Using ReLU activation function) in parallel. with previous two CNN layers. The facial landmarks are obtained by using AAM active appearance model and supervised descent method SDM also named as IntraFac that uses SIFT features. The polynomial learning rate is applied. Various face databases are used to this architecture [18] some of them are CMU MultiPIE with 750,000 images of 337 people under multiple viewpoints, and different illumination conditions. MMI database with 11500 images, CK+ i.e., Extended Cohn-Kanade database containing 309 images.A naturalistic database Denver Intensity of Spontaneous Facial Actions (DISFA) and many more. The method gave higher accuracy top-2 expression for wild datasets i.e., FERA, SFEW, FER2013 inferring to transitions. The proposed architecure couldnt compete with AlexNet for the DISFA and SFEW dataset. With limitations to the setting of pose and lightning.

3. Hypotheses-Restrictions

The predominant limation observed for the [20] [19] that uses Gabor wavelet feature extraction is its High dimension and high redundancy eventhough it has a maximum variance of features that can be obtained. While the segmetation [5] of images has extracted features of the neutral JAFFE dataset images better than Gabor wavelet feature extraction, There has been further research in optimizing the Gabor wavelet method that can be applied to get better features with less dimension and redundancy. The other major observation is the use of grayscale images for the FER to develop and assess the model with datasets as JAFFE and YALE etc., and also performing it for the colored image dataset to get similar satisfactory classification,

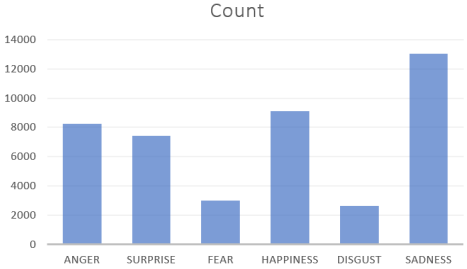
The facial expressions such as contempt, neutral are mild and prone to be having a higher classification error. This is to be overcome by the implementation 2-D Valence and Arousal that has given better performance in the wild [1]. The use of architecture applied in the [18] with inception layer idea concept may be highly fetching. The MRC classifier implemented in the [5] is suitable for simple feature vector data that has been obtained with 2DPCA. However, it works based on euclidean distance and hyperplane for classification cant be well defined. Hence, we can implement a SVM classifier for the feature vectors and obtain better recognition rate. The [5] was somewhat successful in picking the features of the mild emotion in the JAFFE dataset. It can be further tested in the CK+ Extended cohn-Kohn dataset that has similar contempt as mild feature. The feature extraction has always been a bottleneck for the image recognition with different variables such as lightning, localization, environment affecting the recognition rate at the end.

The use of Viola-Jones descriptor is seen in most methods over the literature survey for the face detection. However, with the recent advances in the graphical hardware. There is a room to implement a face detection descriptor algorithm that combines both the Viola-Jones descriptor and the Convolution neural networks offering faster and minimal memory results. The affect recognition using Deep Neural Networks is also in the current trend of face detection, an example of it in the [1]. However, no state of the art performance has been observed between the SVM which has been mostly recommended in ML models for classification and Convolution Neural Networks, methods for the static image datasets.

4. Data Preparation

The custom dataset contains some images from the Aff-Wild2 dataset that can referenced to or used and described in [6] [17] [10] [12] [16] [14] [11] [13] [22] [8] using the methods [15] [7] [9] [21]. and generating the images using techniques described and implemented in the [7] [21] [15] [9]

The dataset contains total of 43435 images of various individuals displaying six emotions which are Anger, Disgust, Sadness, Surprise, Happiness, Fear [3]. The distribution that can shown in the figure below.



The data has been extracted and loaded for Preprocess-

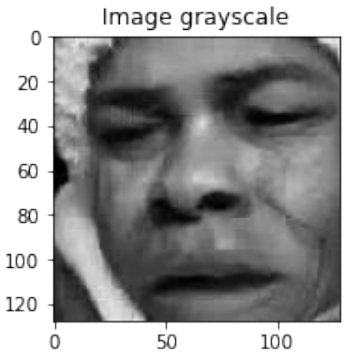
ing. However, before proceeding with that step we need to verify if there are any duplicates in the dataset and remove them. As, there might be chance of having duplicates in sadness emotion with its high distribution and that may lead to imbalanced classification training, Performed the removal of duplicates to the dataset with help of hashing of around 196 images in dataset with the method of hashing. This can further refined and obtain more duplicates in the sadness folder. The duplicated image can seen below.

.

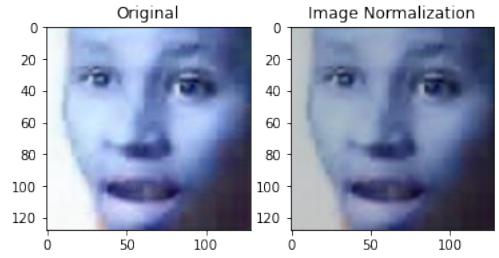
The classes are encoded just in case since they are categorical into numerical for better analysis and model training of classification. Then can perform Oversampling of lower distribution classes implementing some steps seen in the data preprocessing.

5. Data Pre-processing

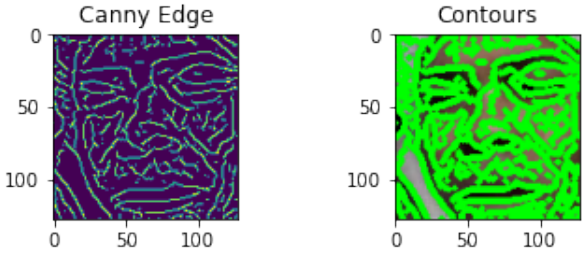
There can be various Data Preprocessing techniques applied to the data images for obtaining more number of quality features such as primarily converting the image in grayscale. That is performed and seen in the below image.



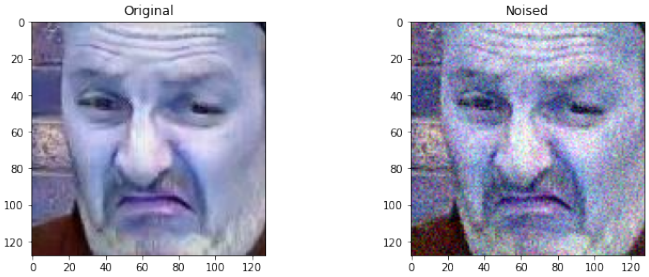
Now further we can apply normalization to images that will further provide contrast to face and its surrounding background with respect to median as seen in the [20]. The image is shown in below.



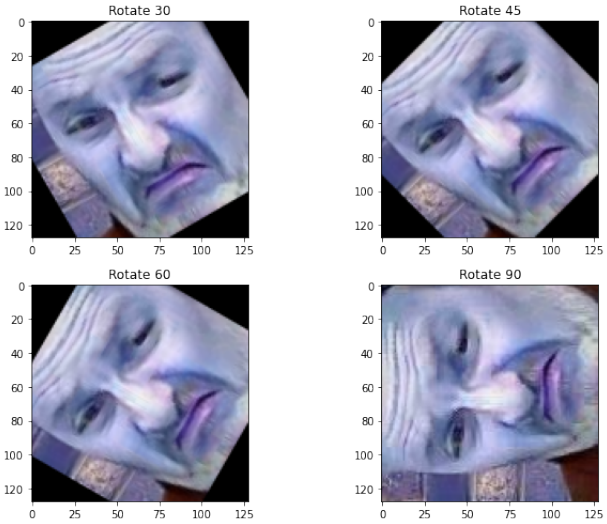
As seen in the [2] to get better mild emotion features we can implement Canny method of Preprocessing. That gives the image with edges defined perfectly to aperture detail we require set to 7 with high detail.



The image beside the canny edge is the contours preprocessed image that has 199 contour points that can be highly beneficial for better feature extraction and training of model. The disgust class has few images data so performed data augmentation that can be done in various ways or techniques. Now currently added noise to increase the robustness and performance of model classification while training.



Then performed the augmentation technique of rotating the image completely and partially to train the model to take the facial action units better while in real, that can seen in the figures below.



References

[1] Estimation of continuous valence and arousal levels from faces in naturalistic conditions. 3. 2, 3

[2] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Facial expression recognition based on facial components detection and hog features. 2014. 2, 4

[3] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999. 1, 3

[4] R.J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice, 2nd edition. *OTexts*, 2018. 1

[5] Dewan Imdadul Islam, S. R. Ngamwal Anal, and Aloke Datta. Facial expression recognition using 2dpca on segmented images. In Siddhartha Bhattacharyya, Nabendu Chaki, Debanjan Konar, Udit Kr. Chakraborty, and Chingtham Tejbanta Singh, editors, *Advanced Computational and Communication Paradigms*, 2018. 3

[6] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 3

[7] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[8] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5):1455–1484, 2020. 3

[9] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017. 3

[10] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 3

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 3

[12] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 3

[13] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 3

[14] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 3

[15] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020. 3

[16] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 3

[17] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 3

[18] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2016. 3

[19] Ebenezer Owusu, Yongzhao Zhan, and Qi Rong Mao. A neural-adaboost based facial expression recognition system. *Expert Systems with Applications*, 41(7):3383–3390, 2014. 1, 2, 3

[20] Ebenezer Owusu, Yonzhao Zhan, and Qi Rong Mao. An svm-adaboost facial expression recognition system. *Applied Intelligence*, 2014. 2, 3, 4

[21] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022. 3

[22] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 3