Survey Paper on Facial Emotion Recognition

Balaji Balasubramanian
Information Technology
Vidyalankar Institute of Technology
Mumbai, India
balajib26@gmail.com

Pranshu Diwan
Information Technology
Vidyalankar Institute of Technology
Mumbai, India
pranshusdiwan@gmail.com

Rajeshwar Nadar
Information Technology
Vidyalankar Institute of Technology
Mumbai, India
rajeshwar.nadar@vit.edu.in

Anuradha Bhatia
Information Technology
Vidyalankar Institute of Technology
Mumbai, India
anuradha.bhatia@vit.edu.in

***Abstract* — Human beings rely a lot on non-verbal communication and facial emotion is a large part of it. In this review paper we cover the datasets and algorithms that are used for Facial Emotion Recognition (FER). The algorithms range from simple Support Vector Machines (SVM) to complex Convolutional Neural Network (CNN). We explain these algorithms through the fundamental research papers and go through their application to the task of FER.**

Keywords—emotion recognition, facial emotion recognition, convolutional neural network, support vector machines.

I. INTRODUCTION

The way of interaction between human beings is more than just verbal communication. According to scientific studies, humans rely a lot on non-verbal methods of communication, specifically communication and understanding each other through facial expressions. While verbal methods of communication may be primarily used, a lot can be derived from a person's non-verbal communication, particularly their expressions.

Facial expressions are more explanatory in situations where words fail, like a shock or a surprise. Moreover, lying through spoken words is more difficult to detect compared to faking expressions. Statistically non-verbal forms of communication make up about 66% of the total communication. This makes it very essential to study, use and analyze facial expressions.

The use of analyzing facial expressions could be put to a variety of use cases. Consider a scenario in which the system detects drowsiness of a driver in the car, and if detected, cause the alarm and stop the car.

According to National Highway Traffic Safety Administration (NHTSA), accidents due to drowsiness cause around 1500 deaths annually [29], and thus makes such systems unavoidable. Implementing such systems will greatly help in reducing the number of accidents and the costs occurred with them.

In another scenario, consider a psychologist using this system to detect what their client might be feeling so that the diagnosis might be done better because in this case, nonverbal communication like body language and facial expressions reveal a lot more than words.

II. DATASETS

We considered five datasets on which we could train our models. A description of each is given as follows:

- CK+ [7]: The CK+ dataset has 593 sequences from 123 subjects. Each sequence has 10 to 60 frames. The dataset has people of diverse ages. Image sequences are frontal views and 30-degree views and are stored as pixel arrays of dimension 640x490 or 640x480 in gray scale or color scale. The images are labelled with seven basic emotion categories: Anger, Contempt, Disgust, Fear, Happy, Sadness and Surprise.
- JAFFE [8]: The JAFFE dataset has 219 images of Japanese females. The images are labeled with Six basic Facial expressions (happiness, sadness, surprise, anger, disgust, fear) and a neutral face.
- FER 2013 [9]: The FER 2013 dataset was created during ICML 2013 Workshop on Challenges in Representation Learning and it contains 35887 48x48 images labeled with seven basic emotion categories which are Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. Google image search API was used to collect the images that were used in this dataset.
- AFEW [10]: Most facial expression datasets have been created in artificial lab environments where the people were asked to generate the required facial expressions. AFEW dataset has short video clips that contain spontaneous facial expression collected from movies. It has seven basic expressions the same as FER-2013. The dataset also captures natural facial expressions of the people with nature head poses, diversity of race and gender. Many clips have multiple people in them which can be used to train the algorithm to capture multiple facial expressions in one clip. The clips include scenes with both indoor and outdoor backgrounds, captured during day and night time to provide different lighting conditions. This can help researchers build algorithms that can generalize better.
- EmotioNet [11]: EmotioNet dataset contains 1 million images downloaded from the internet which are taken in varying natural background and are labelled with 23 basic or compound emotion categories defined in Du et al. [12].

III. Machine Learning Methods

A. *Support Vector Machines (SVM)*

Vapnik introduced SVM in 1992 [33] and it is found to be theoretically and mathematically simpler than Artificial Neural Networks (ANN) [1]. SVM's are supervised learning models used for classification and regression. Given a training set, it outputs an optimal hyperplane which classifies new data points. SVM maximizes the margin between the training patterns and the decision boundary, thus tuning the capacity of the classification function.

For automated real time facial expression recognition, the input data being from a video stream, we can employ an automatic facial feature tracker which performs face localization and extracts 22 feature points [2]. This model can then be used to dynamically classify unseen feature points and return the results to the user.

Optimization of the parameters can be explained using the following tuning parameters: regularization, gamma and the margin. Regularization refers to the margin of the hyperplane; and how much do we want to misclassify each point in the training set. A high regularization parameter tends to classify all the training points correctly. A low regularization parameter tends to misclassify some points in the data, as shown in figure I.

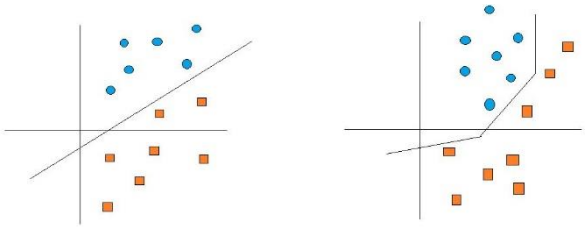

Fig. 1. Regularization parameter in SVM

The gamma parameter decides the distance of the training points from the separating hyperplane. A low gamma means points far away from the separating hyperplane are considered, and a high gamma parameter means only the nearby points are considered. It is illustrated in fig no.

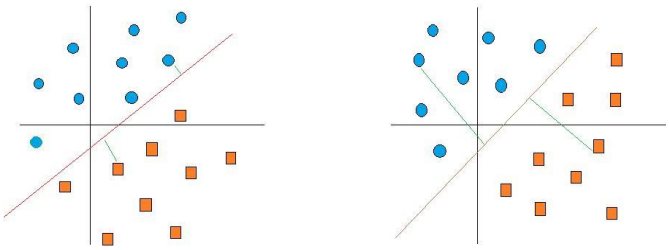

Fig. 2. Gamma parameter in SVM

All SVMs try to achieve a good margin. Margin is a line separating two classes of defined data points. A good margin is as far as possible from both sets of data points. Thus, it makes sense that the margin be equidistant from both sets of points, as shown in the figure.

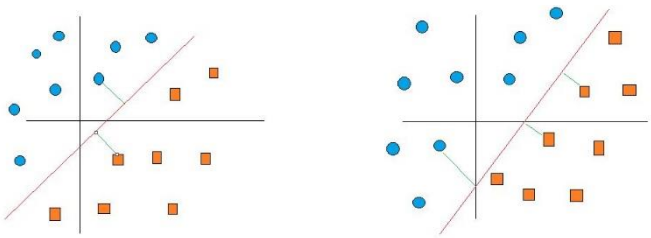

Fig. 3. Margin parameter in SVM

The process of giving a result output emotion by using SVM can be done in the following method. There are 3 core steps involved: a) Face detection in an image, b) Facial data extraction and c) classification of facial emotion [4]. For a real-time video stream, localization is required for detection of human face, after which we crop and grey scale the image, and finally resize it into a 48X48 pixel size as an input to our algorithm.

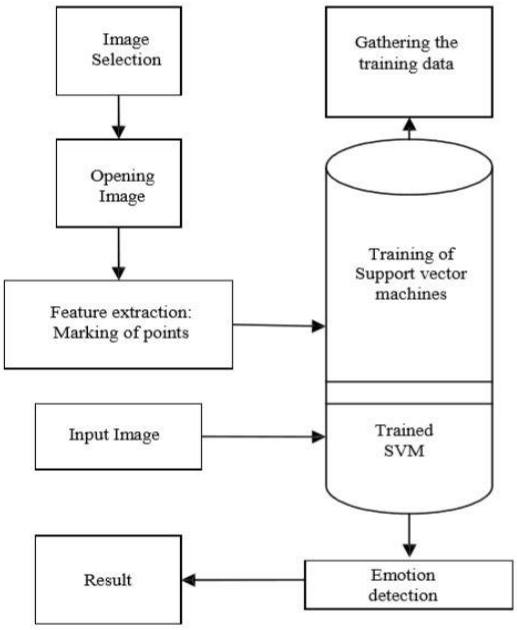

Fig. 4. Flow design of FER using SVM

For real-time automated facial expression recognition using SVM, we initially have a training face dataset, like the FER-2013, along with a real time video capturing device, like a webcam or a video camera. We get our input image by capturing frames from our webcam. The first step involves the detection of the human face inside the captured frame. This image is then vectorized. Meanwhile, SVM gets trained on the training data, shown in the second block and thus the trained model is ready. This model can now be used to classify images and test out results. We test this model out on an image not in the training dataset and display its output.

B. *Using Linear Discriminant Analysis*

For more than two classes (in our case, seven), Linear Discrimination Analysis is the preferred linear classification technique. LDA makes use of the Bayes theorem and makes predictions by estimating the probability of a set of inputs belonging to each class. LDA are more optimized when the

class distributions are gaussian, but in non-gaussian distributions, SVMs outperform LDA [3].

For real-time face detection on the DFAT-504 dataset using Linear Discriminant Analysis, the results are significantly lower than SVM [3]. The performance can be improved by adjusting the threshold for each emotion. But even with threshold adjustment, SVMs outperform LDA.

C. *Using Hidden Markov Models*

A Markov chain is a set of states of a system linked together used in real world computer simulations. For a given Markov chain, we can get the probabilistic simulation of the next state. Using a hidden Markov Model (HMM**)** [5], we can predict the set of unknown variables from a set of known variables. It can give us the probabilities of state transition. HMM's give this probability based only on the current state the model is in, and does not consider the previous states.

Facial Animation Parameters (FAP) describe the Face shape and its movements. [30] FAP's that control eyebrow and mouth movements contain a lot of information about facial expressions. We can build our Markov Models on our defined FAPs to classify them based on six facial expressions [6]. FAPs are extracted from the video and passed to HMM-based FER system and facial expression is recognized.

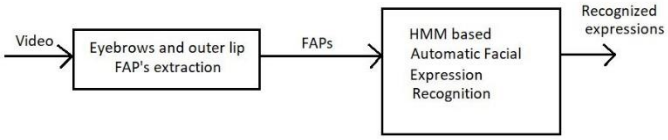

Fig. 5. HMM based FER system

IV. Basic CNN

- Alex Net-Alex et-al [13] trained a Deep Convolutional Neural Network (DCNN) called AlexNet on Imagenet ILSVRC Dataset and achieved a top-5 error rate of 15.3% and won the ImageNet ILSVRC-2012 competition [14]. This made DCNN popular for the task of Image Classification. AlexNet Alex Net has 60 million parameters. It has five Convolutional Layer and three fully connected layers. The Convolutional Layer was followed by Rectified Linear Units [15] to add non-linearity and Max Pooling Layer to reduce the number of Parameters and helps deal with overfitting. The fully connected layers are followed by dropout layers. Dropout [16] is a technique in which few neurons are randomly set to zero. In AlexNet the dropout coefficient is 0.5 i.e. output of the neuron is set to zero with the probability of 0.5. Dropout reduces overfitting and co-adaptation of neurons. They use data augmentation which artificially enlarges the dataset.
- VGG Net [17]- They used small(3 x 3) convolutional filters with stride 1 and managed to push the depth of the Neural Network to 16-19 layers. The Network was able to learn more complex features due to the increased depth and brought down the top-5 error rate in ILSVRC dataset to about 7 percent. It has over 130M parameters and about 90% of those is due to the Fully Connected Layers.
- GoogleNet [18]- It is a 22 layer Neural Network and it won the Imagenet Competition 2014. It uses 12x fewer parameters than Alex Net. It achieved this by omitting the Fully Connected Layers. The authors wanted to create a network that uses fewer computational resources while providing high accuracy because it would reduce power and memory consumption and can be used for many practical purposes.
- ResNet [19]- The Resnet has 152 layers, 8x times deeper than VGGNet and managed to achieve 3.57% top-5 error on ImageNet Dataset. The authors found that when they increased the number of layers by adding Convolution Layers the accuracy did not reduce and got saturated because the Network because too complex and it was hard to optimize using backpropagation. So the authors introduced Resnet Blocks given in the figure below in which H(x)=F(x) + x in which an identity function is added which turns out to be a shortcut during the optimization of the Network. This allowed them to add a lot of layers to the network which would allow it to learn complex features. The network also performs well in the task of Object Detection and Image Segmentation in which the complex features are very helpful.

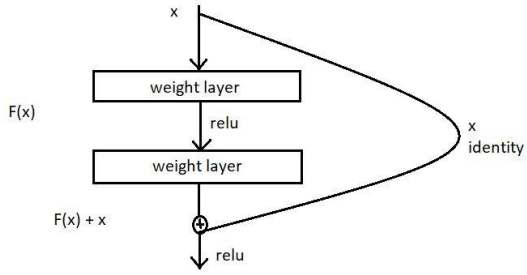

Fig. 6. ResNet layer

V. Deep Learning

In Zhiding et al [20] the authors first detected the face from the image. The dataset contains labeled movie images. There is a lot of background noise, so it is important to first detect the face and then it was fed to Deep CNN to find the emotion. To detect the face an ensemble of face recognition algorithms was used as shown in figure 7. The ensemble consists of joint cascade detection and alignment (JDA) detector [31], the Deep CNN based detector [32] and Mixtures of trees (MOT). The network is first pre-trained on FER dataset and then finetuned on SFEW dataset.

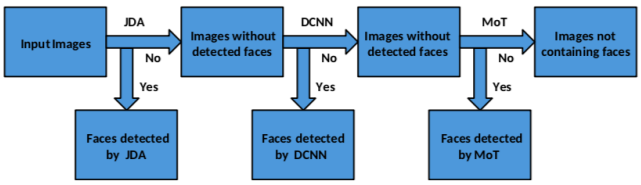

Fig. 7. Detecting faces in an image [20]

The Deep CNN architecture used to detect the emotion is shown in figure 8. It takes in 48X48 dimensional grayscale image as its input and has five convolutional layers, three stochastic pooling layers [21] and 3 fully connected layers. Stochastic pooling layer is used instead of max pooling layer because it gives better accuracy when there is less training data. Max pooling chooses the maximum value whereas stochastic pooling randomly selects the value based on probability distribution obtained by normalizing the values which add randomness thereby improving the network's performance. Fully connected layers contain dropout which is also a mechanism for randomization. This statistical randomness improves the network's generalization and reduces the risk of overfitting.

Random perturbation (data augmentation) to input faces is used which generates additional unseen training samples and improves network performance.

In Octavio et al [22], the authors implemented a convolutional neural network (CNN) in real time in a robot platform which has limited hardware capabilities for Emotion and Gender Classification. They proposed two CNN models. Both models were designed to provide the accuracy to size ratio. The first model is a fully-convolutional neural network and does not contain fully connected layers. It is composed of 9 convolution layers, uses ReLUs [24] as the activation function, Batch Normalization [25] is used to speed up the learning process and Global Average Pooling is used before the softmax layer. This model contains approximately 600,000 parameters. It did not have any fully connected layers because they contain a lot of parameters and increase the size of the model. The last convolutional layer has to replace the fully connected layers so the same number of feature maps as the number of classes. This followed by an average pooling layer which further reduces the number of parameters. The final layer is a softmax layer which to give the output prediction.

The second model is based on Xception [26] architecture. This architecture uses residual modules [27] and depth-wise separable convolutions [28]. Depth-wise separable convolutions contain depth-wise convolutions and pointwise convolutions. It is shown in figure 9. Depth-wise convolutions contain D × D filter which is applied to each input channels separately and point-wise convolution contains 1 × 1 convolution which reduces the number of channels from M to N (M > N). This is done to reduce the number of parameters of the Deep CNN. The final architecture is a fully-convolutional neural network that contains 4 residual depth-wise separable convolutions along with batch normalization and uses ReLU as the activation function. They got an accuracy of 96% IMDB gender dataset and 66% on the FER-2013 emotion dataset. They used guided-gradient back-propagation proposed by Springenberg et al [23] to visualize the features learned by CNN.

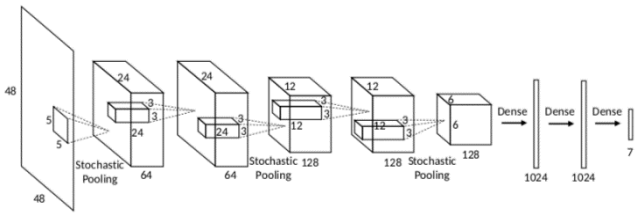

Fig. 8. Deep CNN architecture for detecting FER [20]

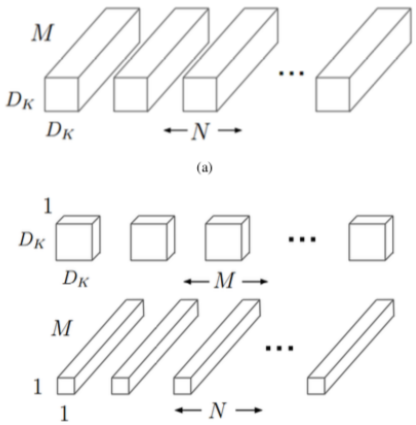

Fig. 9. Depth wise convolution [22]

VI. Conclusion

We covered the datasets and algorithms that are used for the task of FER. The images in the dataset need to be taken from the natural environment so that they close the environment the algorithm is deployed in. We need to gather more training examples for less common emotions like Disgust. CNN outperforms Machine Learning algorithms like SVM. The pipeline of first performing face detection and training a CNN to its output performs well.

Acknowledgment

The authors would like to thank all our anonymous critics for their feedback on this paper, and our project guide, along with the faculty of Vidyalankar Institute of Technology for their unconditional support.

References

[1] Ahmad, A. R., Khalid, M., and Yusof, R. (2002). Machine Learning Using Support Vector Machine. Proceedings of the 2002 Malaysian Science and Technology Congress. 19- 21 September 2002. Johor Bahru, Malaysia. 1-8.

[2] P. Michel and R. El Kaliouby, "Facial Expression Recognition Using Support Vector Machines," 2000.

[3] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '05), pp. 568-573, 2005.

[4] Bajpai A, Chadha K: Real-time facial emotion detection using support vector machines. *Int. J. Adv. Comput. Sci. Appl* 2010, 1(2):137-140.

[5] NEFIAN, A. V. AND HAYES III, M. H. 1998. Hidden Markov models for face recognition. In Proceedings, International Conference on Acoustics, Speech and Signal Processing. 2721–2724.

[6] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream hmms," TIFS, vol. 1, no. 1, pp. 3–11, 2006.

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

[8] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference*

[9] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[10] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[11] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA*, 2016.

[12] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[13] Alex Krizhevsky, Sutskever I, and Hinton G.E, Imagenet classification with deep convolutional neural networks. In NIPS, 2012

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009

[15] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. 27th International Conference on Machine Learning*, 2010.

[16] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, Going deeper with convolutions, CoRR, 2014.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[20] Zhiding Yu, Cha Zhang Image based Static Facial Expression Recognition with Multiple Deep Network Learning, 2015

[21] M. D. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557, 2013.

[22] Octavio Arriaga, Paul G. Plo¨ger and Matias Valdenegro. Real-time Convolutional Neural Networks for Emotion and Gender Classification 2017

[23] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[24] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[26] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017

[29] Assari MA, Rahmati M. (2011). Driver drowsiness detection using face expression recognition. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 337– 341.

[30] I. S. Pandzic and R. Forchheimer, Eds., *MPEG-4 Facial Animation*. New York: Wiley, 2002.

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2879-2886, IEEE 2012.

[32] C. Zhang and Z. Zhang, improving multiview face detection with multi-task deep convolutional neural networks, in Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, pages 1036 – 1041, IEEE, 2014.

[33] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A Training Algorithm for Optimal Margin Classifiers In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA, 1992. ACM Press.