Facial Expression Recognition using Preprocessing and Hybrid network

Vivek Sohal[1], Ajinkya Pawale[2], Nikhil Dalvi[3], Roshan Talreja[4], Sunita Sahu[5]
[1,2,3,4,5] Dept. of Computer Engineering, V.E.S.I.T, Mumbai, India.
[1]2015vivek.sohal@ves.ac.in, [2]2015ajinkya.pawale@ves.ac.in, [3]2015nikhil.dalvi@ves.ac.in, [4]2015roshan.talreja@ves.ac.in
[5]sunita.sahu@ves.ac.in

Abstract: Facial expression recognition is rapidly becoming an area of intense interest in computer science and human-computer interaction design communities. The most expressive way humans display emotions is through their facial expressions. The core module of our emotion recognition system is a hybrid network that combines recurrent neural network (RNN) and 3D convolution network (C3D). Recently, fully-connected and convolution neural networks have been trained to achieve state-of-the-art performance on a wide variety of tasks such as speech recognition, image classification, natural language processing, and bioinformatics. For classification tasks, most deep learning models employ the softmax activation function for prediction which minimizes cross-entropy loss. In this paper, we propose a small but consistent advantage of replacing the softmax layer with Support Vector Machine in the hybrid network. This approach minimizes a margin-based loss instead of the cross-entropy loss. Also, preprocessing steps such as synthetic sample generation, rotation correction, cropping, down-sampling, and intensity normalization play a major role to train samples that have limited data set. The purpose of this system is to evaluate the facial expression and classify their expressions into one of the following emotions: angry, disgust, fear, happy, neutral, sad and surprise.

Keywords - Face Expression Recognition, Preprocessing, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), 3D Convolution Neural Network (C3D), Hybrid CNN-RNN and C3D Network, Support Vector Machine (SVM).

I. INTRODUCTION

Facial expression recognition is an important topic in the fields of computer vision and artificial intelligence owing to its significant academic and commercial potential. It is an area where a lot has been done and a lot more can be done. With great progress in intelligent systems in recent years, expression recognition remains the most important problem for human interaction. Expression recognition is challenging due to the difficulties for definition and classification of emotion expressions for different people without contextual or psychological information. Facial expression recognition is not a theoretical field but finds practical applications in many fields. Coupled with human psychology and neuroscience it can come up as an area which can bridge the divide between the more abstract area of psychology and the more crisp area of computation. Its associated research is minherently a multidisciplinary enterprise involving a wide variety of related fields, including computer vision, speech analysis, linguistics, cognitive psychology, robotics and learning theory, etc. A computer with more powerful expression recognition intelligence will be able to better understand human and interact more naturally. Many real-world applications such as commercial call center and affect-aware game development also benefit from such intelligence. Possible sources of input for expression recognition include different types of signals, such as visual signals (image/video), audio, text and bio signals. For vision-based expression recognition, a number of visual cues such as human pose, action and scene context can provide useful information. Nevertheless, the facial expression is arguably the most important visual cue for analyzing the underlying human emotions.

This paper is further organized as follows: Section II provides an overview of related work on Facial Expression Recognition. Proposed Architecture is discussed in Section III. Methodology of the system is explained in Section IV. Expected results are presented in Section V, leading to conclusions in Section VI.

II. RELATED WORK

In [1], authors Manglik et al. proposed two phases in Facial Expression Recognition. The first phase is image processing and the second phase is setting and training of the neural network. The image processing phase or the preprocessing phase involves five steps. In the first step, the image of the face is normalized. The normalized image is subjected to a grayscale transformation in the second step. The third step is performed by partitioning of the transformed image in two portions: the upper half and the lower half. In the fourth step, a frequency analysis of the normalized image is performed in the upper half partition. This tracks the position of the eyes and the eyebrows. Contouring is performed to trace the shape of the eyes and the eyebrows. In the fifth step of the first phase, a further frequency analysis in the lower half reveals information about the nose, the mouth, and the cheeks. The contours are vectored to obtain a feature vector. In the next phase, this feature vector is used for setting and training of the Hopfield Neural Network. The advantage of this method is that it is computationally inexpensive. The method is the one which appears more natural and a bit easier to understand as the authors tried to relate it with basic mathematics and signal theory. The limitation of this method is that there is only four emotion classification: happy, angry, sad and surprise against neutral. Also, for tracking the position of the eyes, the assumption is that the face should be upright (Up to 30 degrees is acceptable) and the eyes lie in the upper portion of the face.

Authors Sarode et al. implemented a method using 2D appearance-based local approach for the extraction of intransient facial features and recognition of four facial expressions: happy, angry, sad and surprise against neutral as per [2]. The algorithm implements Radial Symmetry Transform and further uses edge projection analysis for feature extraction and creates a dynamic spatio-temporal representation of the face, followed by classification into one of the expression classes. This algorithm achieves an accuracy of 81.0% for facial expression recognition from the grayscale image. This method describes a more robust system for facial expression recognition from static images using 2D appearance-based local approach for the extraction of intransient facial features, i.e. features such as eyebrows, lips, or mouth, which are always present in the image. The advantage of this method is that it has low computational requirements. The limitation of this method is that there is only four emotion classification: happy, angry, sad and surprise against neutral. Also, frontal view of the image should be available for classification.

For classification tasks, most deep learning models employ the softmax activation function for prediction and minimize cross-entropy loss as per authors Tang et al. in [3]. It demonstrates a small but consistent advantage of replacing the softmax layer with a linear Support Vector Machine. This approach minimizes a margin-based loss instead of the cross-entropy loss. While there have been various combinations of neural nets and SVMs in the prior art, results of this method using L2-SVMs show that by simply replacing softmax with linear SVMs gives significant gains on different datasets. The advantage of this method is that there is a consistent performance gain in the system by replacing the softmax layer with linear SVM. The limitation of this method is that it does not use multiclass SVMs instead of linear SVMs.

In [4], authors Yu et al. seek to automatically classify a set of static images into 7 basic emotions: anger, disgust, fear, happy, neutral, sad and surprise. The proposed method contains a face detection module based on the ensemble of three state-of-the-art face detectors (JDA, DCNN, and MOT), followed by a classification module with the ensemble of multiple deep convolution neural networks (CNN). Each CNN model is initialized randomly and pre-trained on a larger dataset Facial Expression Recognition (FER) 2013. These pre-trained models are then one-tuned on the training set of SFEW 2.0. Fine-tuning proved to be crucial in boosting the classification performance on SFEW. To combine multiple CNN models, it presents two schemes for learning the ensemble weights of the network responses: by minimizing the log likelihood loss, and by minimizing the hinge loss. This method generates the state-of-the-art result on the FER dataset. It achieves 55.96% and 61.29% respectively on the validation and test set of SFEW 2.0. The main contribution of this method is 1. To propose a CNN architecture that achieves excellent emotion recognition performance. 2. To propose a data perturbation and voting method that further increases the recognition performance of CNN considerably. 3. To propose two novel constrained optimization frameworks to automatically learn the network ensemble weights by minimizing the loss of ensembled network output responses. The limitation of this method is that for any frame with multiple face detections, the largest face is returned. This strategy mostly works well except in very occasional cases where the largest face is not intended for emotion recognition.

The core module of [5] proposed by authors Fan et al. is a hybrid network that combines recurrent neural network (RNN) and 3D convolution networks (C3D). RNN and C3D models encode appearance and motion information in different ways. Specifically, RNN takes appearance features extracted by convolution neural network (CNN) over individual video frames as input and encodes motion later, while C3D models appearance and motion of video simultaneously. This method achieved a recognition accuracy of 59.02% without using any additional emotion-labeled video clips in the training set. Extensive experiments of this method show that combining RNN and C3D together can improve video-based emotion recognition noticeably. The main contribution of this method is that the hybrid network of LSTM and C3D gives impressive results. In C3D, the output softmax layer can be replaced by SVM to further improve the performance of the system. Also, the number of C3D models can be increased to improve the performance of the system.

Authors Lopez et al. use a combination of CNN and preprocessing steps as per [6]. The model is fast to train and it uses real-time facial recognition. The paper also mentions how training and testing environment plays a major role in facial expression recognition. Different ethnic groups, different conditions, etc. affect the accuracy of the system and they give a better idea about the performance of the system in these conditions. So, it is important to perform cross-database evaluation i.e. training on one database and testing on other. This paper uses static based images with output as 7 expression set (6 basic + neutral) for the controlled and uncontrolled environment. This model has alternating types of layers, including convolution layers (Kernel size and generated maps), sub-sampling layers (Increase in positional variance by reducing map size) and fully connected layers. It uses a gradient descent method in CNN for effective results and better learning. The main steps of this model are image normalization, synthetic training sample generation, and CNN. Training time is small as compared to other systems and output is given in real-time. The main thing is the usage of training, validation and testing sets. Order change using gradient descent increases the accuracy of the system. For this, the validation set selects the best possible order. The main advantage of this paper is the use of preprocessing steps. Synthetic data generation (For large sample generation which is important for CNN), rotation correction (Eliminate factors that negatively affect accuracy), cropping (To focus on expression specific features), down-sampling (Reduce pixel dimensions for better performance), intensity normalization (Brightness and contrast are countered). Following are the limitations that can be taken into account and improved. This paper requires the proper location of each eye for image processing. Also, the accuracy of certain expression (for e.g. sad) is less as compared to others. Frontal images with the controlled environment are needed for better accuracy.

In [7], authors Mollahosseini et al. proposed a method in which two convolution layers followed by max pooling and four inception layers are used. Results are comparable or better than the normal state of the art methods and traditional convolution neural networks in both accuracy and training time. This technique allows cross-database classification while training on databases with limited scope. Deep neural networks are suitable for the cross-database as it performs well in flexible learning tasks as compared to hand-coded features. As the depth and topology of neural networks are increased, it leads to over-fitting of training data and increased computational needs. This can be solved by the deep sparse network but current GPU and CPU systems are not sufficient enough to work with it. Thus, inception layers are used that provides an approximation of the sparse network. Also, the learning rate used is polynomial which is better than exponential used in other places. By using polynomial learning rate, the test loss is converged at a faster rate and it also allows network for many iterations without fine-tuning. The accuracy is less in this paper for some subject independent tasks due to pose and illumination variations. Also, as they have not used linear Support Vector Machines or engineered features used in traditional methods which are better for subject independent tasks.

III. PROPOSED ARCHITECTURE

The proposed architecture is a combination of the following techniques: preprocessing (To cope up with minimum available data), RNN (Because CNN requires a large set of labeled data for training which is difficult for real world application), LSTM (To solve the vanishing gradient problem), C3D (It is better than 2D CNN), hybrid of RNN-C3D (To further improve performance of the system) and SVM (It is an extremely efficient classifier).

A. Preprocessing

For Neural Networks, a fully extensive database is not available for deep architecture. Thus, preprocessing is used to extract specific features from an image and explore presentation order of samples during training as per [6]. The preprocessing steps used are as follows:

a. Synthetic Image Generation: As deep neural networks can take care of distorted images that have imperfection in the eye detection procedure, but it needs a lot of samples which are not available in public datasets. So, synthetic images are used to achieve this task. 30-70 images for each image are synthesized (New eye position equivalent to original one but distributed by a Gaussian noise).

b. Rotation Correction: Rotation and translation in images are not related to the facial expression but are used instead to eliminate factors that negatively affect accuracy. Rotation of synthetic images will generate images which will have a lot of variation as synthetic images have position disturbed by random Gaussian noise.

c. Cropping: It focuses on expression specific parts by the elimination of background, ears, part of the forehead, etc.

d. Down-sampling: Size of the image is reduced to a 32x32 pixel to increase GPU performance. It helps to identify which regions are related to each expression.

e. Intensity Normalization: Variation due to brightness and contrast is handled using intensity normalization.

The preprocessing steps are explained in figure 1.0 below.

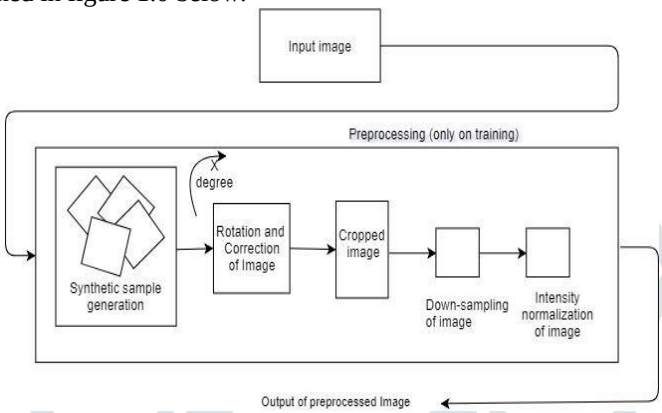

Fig. 1.0: Preprocessing

B. Recurrent Neural Network

Recurrent Neural Network (RNN) is a type of deep learning model. This is a neural network that processes sequential data and takes in as input both the new input at the current timestep and the output (or a hidden layer) of the network in the previous timestep. There is a backward connection between hidden layers in Recurrent Neural Networks. Therefore, they have some kind of memory in them. One can use RNNs in different problems like time series analysis, natural language processing, and speech recognition. Long short-term memory model (LSTM) is one of the most popular, modern RNN. One can construct an LSTM which has a cell state at each timestep that changes with new input using the Python Theano package. Recurrent Neural Network (RNN) is a type of Artificial Neural Network(ANN) where connections between neuron units form a directed cycle. This technique creates an internal state of the network which allows it to illustrate dynamic temporal behavior. RNN is explained in figure 2.0 below.

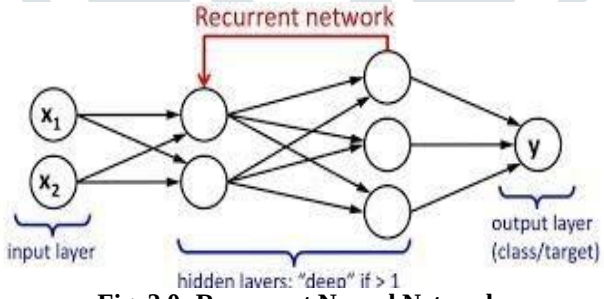

Fig. 2.0: Recurrent Neural Network

C. Long Short-Term Memory

Long Short-Term Memory (LSTM) Neural Networks are a particular type of recurrent neural networks that got a lot of attention recently within the machine learning community. In simpler terms, LSTM networks have some internal contextual state cells that act as long-term or short-term memory cells. The core of LSTM networks is that it can remember a value for an arbitrary length of time. An LSTM unit is equipped with gates that determine when the input is significant enough to remember, when the unit should continue to remember or forget the value, and when the unit should output the value as per [5]. The output of the LSTM network is determined by the state of these cells. This is a very important property when we need the prediction of the neural network to depend on the historical context of inputs, rather than only on the very last input. LSTM is explained in figure 3.0 below.

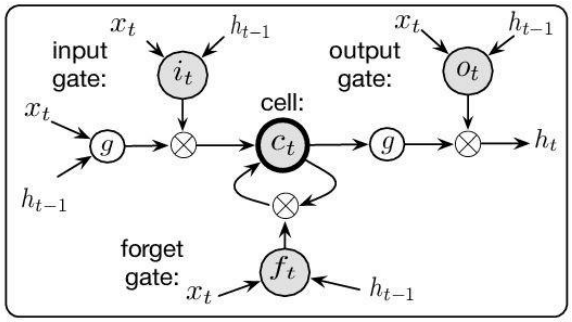

Fig. 3.0: Long Short Term Memory

D. 3D Convolution Neural Network

3D Convolution Neural Network (C3D) is a combined hardware/software process that captures a pair of two-dimensional images, objects, or scenes and automatically reconstructs them into a digital three-dimensional (3D) model. C3D can be understood as a 3D convolution on three channels. In 2D CNN, convolution and pooling operations are only spatially applied to 2D static images. While in 3D CNN, the operations are performed spatio-temporally by adding an additional time dimension. Hence, such C3D networks preserve the temporal information of the input signals, resulting in a more distinctive result as per [5]. Capturing images using the C3D process allows the image to be manipulated horizontally, vertically, or spatially using one of three modeling representations: a polygon wire-frame, shaded or naturally rendering. A current C3D application includes tire footprints, health care studies, police mugshots, virtual studio applications for TV, microscopic modeling objects, aerial road surveys, and museum artifact archiving. C3D is explained in figure 4.0.

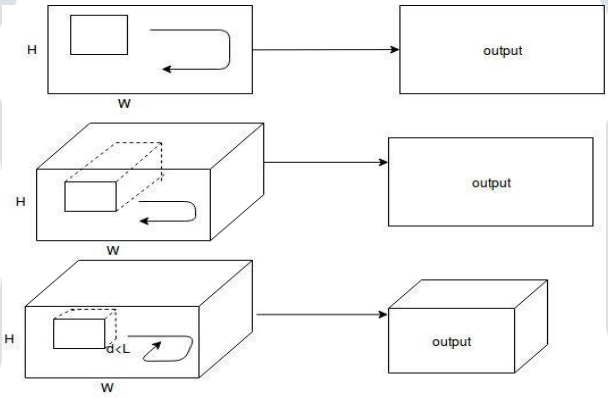

Fig. 4.0: 3D Convolution Neural Network

E. Hybrid CNN-RNN and C3D Network

Either CNN-RNN or C3D can alone achieve good performance. But results according to [5] show that combining RNN and C3D to form hybrid CNN-RNN and C3D network can further improve the performance of the system.

F. Support Vector Machine

Most deep learning methods for classi□cation using fully connected layers and convolutional layers have used softmax layer objective to learn the lower level parameters.
[3] uses linear SVM's objective to train deep neural nets for classi□cation. Lower layer weights are learned by backpropagating the gradients from the top layer linear SVM as per [3]. Also, the results of [3] show that SVMs are better than softmax output layer.

IV. METHODOLOGY

The detailed methodology for proposed expression recognition model is as follows:
1. An input image is given to the preprocessing model.
2. The input image is preprocessed using synthetic sample generation, rotation correction of the image, cropping the image, down-sampling of image and intensity normalization of the image. Preprocessing is only done before the training model as it helps to cope with a few data samples.
3. Input images along with their respective labels are given as input to the training model. Training model is a combination of CNN-RNN and C3D networks.
4. Validation is done on the trained images by performing order change using Gradient Descent. It helps to find the best network weights. It tries to minimize the error between the actual label associated with an image and the output label predicted by the model. After validation, the best network weights are given as input to the testing model.
5. Testing model is also a combination of CNN-RNN and C3D networks. Both the models predict their confidence level respectively.
6. The better confidence level of the two networks is selected as the final confidence level.
7. The output label associated with the confidence level is returned as the emotion evaluated.

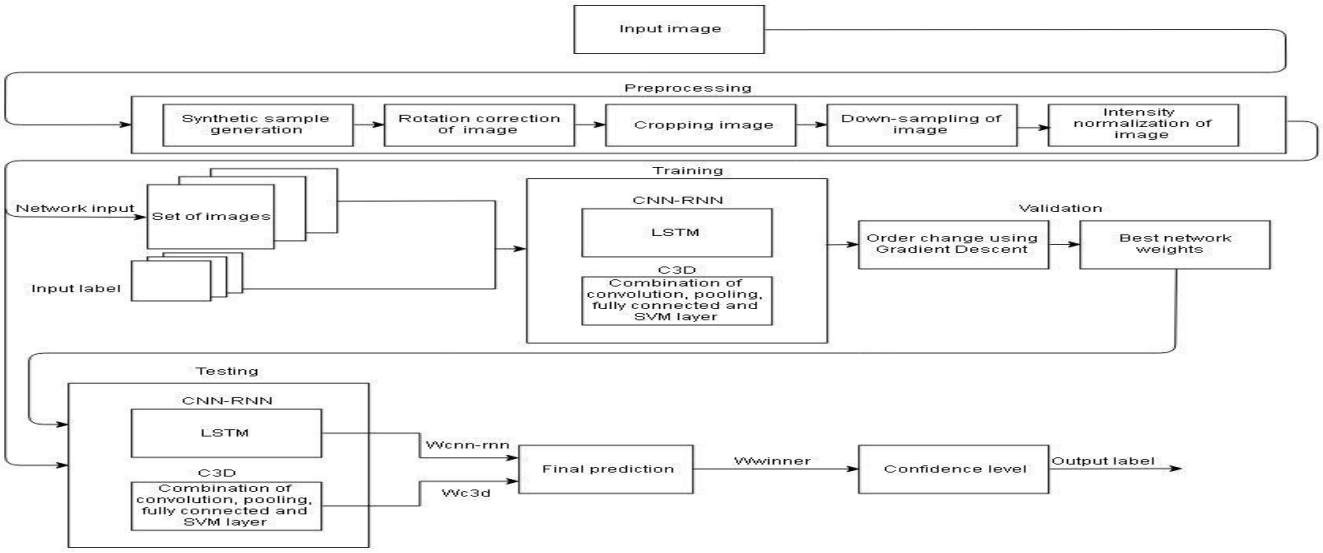

Fig. 5.0: Overview of system

V. EXPECTED RESULTS AND DISCUSSION

Preprocessing steps such as synthetic sample generation, rotation correction, cropping image, down-sampling, and intensity normalization play a major role to train samples that have limited data set and produces better results according to table 1.0 as per [6]. Their accuracy was calculated on CK+ dataset by using individual preprocessing steps as follows:

Table 1.0: Comparison of preprocessing and accuracy of the system

Preprocessing	Accuracy
No preprocessing	53.57%
Image cropping only	71.67%
Rotation correction only	61.55%
Spatial normalization (rotation, cropping and down- sampling)	87.86%
Intensity normalization only	57%
Intensity and spatial normalization	86.67%
Spatial normalization and synthetic samples	89.11%
Spatial normalization, intensity normalization and synthetic samples	89.79%

Results according to [5] show that combining RNN and C3D can improve the performance of the emotion detection system. This system achieved a recognition accuracy of 59.02% using cross dataset evaluation as per [5]. Their accuracy was obtained by training on FER2013 dataset and testing on AFEW 6.0 dataset. The cross dataset experiments show that the proposed approach also works in unknown environments, where the testing image acquisition conditions and subjects vary from the training images. Switching from softmax to SVMs is incredibly simple and appears to be useful for classi□cation tasks according to [3]. Hence, our system based on preprocessing and hybrid network having SVM output layer is expected to achieve a better recognition accuracy than [5].

Application areas for expression recognition: One use case of this system is student behavior analysis. In this, the teacher will get statistics about the class and individual student behavior over course of the lecture. It also gives statistics for students requiring attention. Another use case is marketing/advertising in which expression of customers is recorded. It helps to analyze how people react to an advertisement, product, packaging, and store design.

VI. CONCLUSION

Facial expression recognition has numerous applications in image processing domains, security applications domain or any type of biometric system. This research work is a primary step of expression analysis to classify expressions into one of the following emotions: angry, disgust, fear, happy, neutral, sad and surprise. Preprocessing steps such as synthetic sample generation, rotation correction, cropping image, down-sampling, and intensity normalization play a major role to train samples that have limited data set and less variation. Also, hybrid networks are efficient and help a lot during cross-database evaluation thus improving the accuracy of the system. Replacing softmax with SVMs as the output classifier further increases the accuracy of the system.

REFERENCES

[1] Manglik, Prashant Kumar, Ujjawal Misra, and H. Bindu Maringanti. "Facial expression recognition." Systems, Man and Cybernetics, 2004 IEEE International Conference on. Vol. 3. IEEE, 2004.

[2] Sarode, Neeta, and Shalini Bhatia. "Facial expression recognition." International Journal on computer science and Engineering 2.5 (2010): 1552-1557.

[3] Tang, Yichuan. "Deep learning using linear support vector machines."arXiv preprint arXiv:1306.0239 (2013).

[4] Yu, Zhiding, and Cha Zhang. "Image based static facial expression recognition with multiple deep network learning." Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015.

[5] Fan, Yin, et al. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks." Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016.

[6] Lopes, André Teixeira, et al. "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order." Pattern Recognition 61 (2017): 610-628.

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