A Smart Embedded System to Map Video Feeds of Human Actions Through the Virtual Characters

S. Mushtaq [1,] S. A. Irtaza [2]

[1,2] *Computer Science Department, University of Engineering and Technology, Taxila, Pakistan*

[1]sana.mushtaq@students.uettaxila.edu.pk

Abstract- Virtual reality has made significant advancements in recent years and has the potential to transform the way we interact with digital content. The main focus of this paper is to explore mapping of human actions into virtual characters. One area of research that has gained attention is the integration of facial expression recognition (FER) technology. FER technology involves detecting and analyzing human facial expressions, which can provide important information about a user's emotional state and level of engagement with VR content. The integration of FER into VR can lead to more immersive and engaging experiences, as the technology can allow for more natural interactions between users and virtual environments. Overall, the integration of FER technology into VR holds promise for enhancing user experiences and improving our understanding of human emotions and behavior in virtual environments. This work integrates facial expression recognition with human action recognition (HAR) capable of recognizing human actions. In this paper, we have presented a facial expression recognition mapping from videos to virtual characters. Later on, we have discussed how we can use HAR to recognize human actions. Virtual characters are implemented as 3D models that are results of mapped human actions.

Keywords- Facial Expression Recognition, EfficientNetB7, AffectNet Dataset, RMSProp

I. INTRODUCTION

Facial expressions play a crucial role in human communication by conveying the intentions of others. With the advancement in computer vision tasks and the use of Deep Neural Networks (DNNs), Facial Expression Recognition (FER) has shown significant progress. When provided with only a face, identifying emotions becomes possible through facial expressions, which can be categorized into five major types: happiness, sadness, neutral, surprise, and anger. In recent years, facial recognition has gained attention and has significant potential for both academic and commercial purposes. Numerous studies have been conducted to identify the five fundamental emotional expressions. However, these systems face certain challenges, such as low accuracy on certain datasets such as AffectNet.

Researchers have suggested employing machine learning and deep learning techniques. [1] introduced affectnet dataset that is an improved facial expression recognition dataset that assign many emotion labels instead of single point. The author discussed intra-class variations, inter-class similarities and label noises that appear in existing datasets. The author introduced a new labelling method using binary classifiers. The overall accuracy achieved was 52%. [2] proposed a custom light based CNN model based on MobileNetV2 and aims to provide facial expression recognition with less computational costs. The model achieved 54% accuracy on the affectnet dataset. [3] explores FER using transfer learning on ResNeXt with the affectnet dataset. The paper mentions class imbalances on the dataset and need for better model generation. The study introduced pairwise learning where emotions are classified in pairs rather than all categories simultaneously. It improved minority class recognition. The overall accuracy achieved on the model was 79%. [4] proposes a model called visual transformers with attention feature fusion. The model integrates local binary patterns and convolution neural networks to improve FER in an uncontrolled environment The module combines global and local attention mechanisms with transformer encoder to capture relationships among visual features. The model achieved 61% accuracy. [5] are using 12 layer CNN to detect emotions from facial images. The author discussed the key challenges such as age, cultural variations, head poses and occlusions. The model achieved an accuracy of 86%. [6] introduced a multi stage hierarchical attention network that aims to improve emotion recognition using feature pyramids and fully connected layers to improve facial representation. The model achieved the overall accuracy of 67% on affectnet dataset. [7] proposed a multiscale facial recognition model to enhance interclass separability and intra class compactness in facial expression recognition. The model integrates

dynamic global and static local attention that uses contextual information to distinguish expressions. Deep smooth feature loss that encourages intraclass features. And a multiscale classifier that captures high frequency and low frequency information. The model achieved an accuracy of 63%. [8] trained the Xception algorithm with the imbalanced data distribution with ratio of 99 to 1 on the training and testing datasets. As a result, the validation accuracy was 75.0%. For a facial expression recognition system with noisy annotations. [9] proposed a framework called CCT for effective training of 3 networks. The individual networks in CCT are ResNet-18. The validation accuracy of this model against AffectNet dataset was 66.0%. In [10] a Pyramid crOss-fuSion TransformER network (POSTER) that is based on two stream networks landmark feature and image feature. Landmark features help to detect facial landmark detector while image features work on Convolution Neural Network. The accuracy achieved on AffectNet dataset was 67.3%. [11] mainly worked on the conventional Spatial Transformers to improve its accuracy and the effective attentional regions are captured on facial landmarks or also capture the facial visual saliency maps. They included a module for generating images called masks. Those masked images are fed as input to the STN model. The validation accuracy achieved from this technique was 68.6%. [12] proposed a deep learning approach called Deep Facial Expression Vector Extractor (DeepFEVER) that learns a visual and extract those features to be applied to any FER task or any dataset. It is a convolutional neural network trained on facial expression recognition datasets and an additional unlabeled dataset. The model achieved an accuracy of 65.4%. [13] worked on the approach that helps to combine the features learned by Convolutional Neural Networks (CNN). This approach also helps to achieve the stat- of- art results (FER), a handcrafted feature computed by the bag-of-visual-words model. They experimented with multiple convolution neural networks to obtain automatic features in combination with pretrained models. A k-nearest neighbors' model is applied to select the nearest training samples for an input test image. Then Support vector machines (SVM) classifier is trained on the selected training samples and used to predict the class label for the test image. According to [14] the accuracy on AffectNet dataset turned out to be 63.3% that is based on the Meta-Face2Exp. This framework works to leans expression knowledge on FER data where the adaptation network is trained to fit the labels of data that is generated by the base model. This model uses circuit feedback mechanism that helps to improve the base network with the feedback from the adaptation network.

The test accuracy they achieved on the AffectNet dataset was 64.2%. [15] implemented a geometric model to simulate occlusion for VR headset on existing FER datasets. They used a transfer learning approach on two pertained VGG and ResNet networks. They fine-tuned model on different datasets. The test accuracy on AffectNet dataset by using VGG- Face (transfer learning), ResNet-50 and VGG-Face 16 were 50.1%, 47.3% and 49.2%. [16] worked on AffectNet dataset and proposed a CNN structure by using SIFT features. SIFT descriptor is used to extract the facial expressions features based on 68 landmark points on a face and a CNN is applied over SIFT features. The average accuracy rate achieved from this experiment was 72%.

By achieving improved accuracy on the AffectNet dataset, we used the EfficientNetB7 model that works effectively on unseen data which reduces computational costs, and optimizes data with higher accuracy rate. This paper makes a contribution by combining HAR with FER. Using human action recognition dataset for HAR. Translating human behaviours into virtual characters into corresponding animations in real time. This paper makes a significant contribution based on the following:

- We present an efficient approach for facial expressions recognition on the AffectNet dataset.
- We present an efficient approach for human action recognition on the HAR dataset.
- We present a solution to mapping of real time video feeds on virtual characters.

II. PROPOSED METHODOLOGY

This section elaborates the design procedure and steps involved in developing the model. The proposed methodology for this research work is as follows:

- Action mapping: Facial expressions and behaviour are detected and mapped into characters.
- Response generation: Detected behaviour is processed using EfficientNetB7 and CNN-LSTM. The characters respond with respect to interaction.
- The system utilized a tensorflow based application connected to a device through a camera. The camera captures video feeds and processes to detect actions that control character's animations.

III. DATASET PREPARATION

The initial stage involves preprocessing on a dataset to make it compatible for training and validation purposes. For this purpose, we used an Image Data Generator. We are also doing data augmentation in the same stage as it helps the model to generalize better and avoid overfitting. For FER, we increased data size by applying random transformation over a dataset such as random rotations, horizontal and vertical shifts, shearing, zooming, and flipping of the images. For HAR, Human action recognition dataset is used that contains various labelled human actions. The image

is resized into 224 x 224 x 3.
Where C indicates the number of channels for an image. The settings used for dataset augmentation and preprocessing tasks are displayed in Table 1.

Table. 1 Data Preprocessing & Augmentation Settings

Settings	Value
Rotation Range	20
Width Shift Range	0.1
Height Shift Range	0.1
Shear Range	0.1
Zoom Range	0.1
Horizontal Range	True
Brightness Range	[0.8, 1.2]
Fill Mode	nearest

IV. MODEL ARCHITECTURE

Convolutional Neural Network CNN is a deep learning architecture that is mostly used for image and video processing (Gu et al., 2018). It can detect important features and patterns. In this paper, we adopt EfficientNetB7 as our backbone networks architecture to perform facial expression recognition tasks on the AffectNet dataset. EfficientNet is a CNN based architecture. It is also a scaling method that scales all dimensions of data uniformly by using a compound coefficient φ. The dimensions include depth, width, or resolution with a set of fixed scaling coefficients φ where d is the depth, wi is width and r is the resolution of the image.
The compound coefficient scaling method means that if the input image is large, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image (Tan et al. 2019).
For HAR tasks, we used a CNN-LSTM architecture, where the CNN extracts features from image frames, and the LSTM captures dependencies across sequences of frames. This approach ensures the recognition of different human actions.
The base network EfficientNetB0 network is trained on more than a million images from the ImageNet database. All other EfficientNet models from EfficientNetB1 to EfficientNetB7 are scaled up from base network i.e., EfficientNetB0. However, EfficientNetB7 has state of the art high accuracy on ImageNet. In other words, EfficientNet is small as compared to other models that achieve the same accuracy on ImageNet. For example, the ResNet-50 model has a total of 25.6M parameters but it performs less as compared to the smallest EfficientNet, which only takes a total of 5.3M parameters. The architecture of EfficientNetB7 is as follows:
The EfficientNetB7 architecture is based on a combination of several techniques which comprises compound scaling method to calculate depth width and resolution optimization of the network. This architecture also includes a novel black design called Mobile Inverted Residual BottleNeck (MBConv). The MBConv block is a building block that is used in the EfficientNet family of neural network architecture. It is well suited for the deployment of mobile devices and embedded devices by the use of less computational resources. The block consists of different layers which includes a depthwise convolution layer, followed by a pointwise convolution layer with a bottleneck feature that reduces the number of input channels before expanding them again.
Depthwise convolution layers apply a separate convolutional filter to each input channel reducing computational cost of the operation compared to a standard convolutional layer. It applies a single filter to all input channels. The output of the depthwise convolution layer is passed through the pointwise convolution layer that applies 1x1 convolution filter to the input. It reduces the number of input channels before expanding them again. This helps to reduce the computational cost of the operation and prevent overfitting. The output of the convolution layer is added to the original input to the block, creating a skip connection. The output of the skip connection is passed through an activation function.
The architecture explanation of EfficientNetB7 is as follows:
Input: The input to EfficientNetB7 is an image with a resolution of W x H pixels, which is resized to a smaller resolution before being processed by the network.
Stem: A first layer called convolutional layer with 64 filters and kernel size of 3x3. This is followed by a batch normalization layer and an activation function. The stem is designed to extract the low-level features from the input image.
Blocks: Each block contains several MBConv blocks. Each block is examined by a compound scaling method that helps to scale depth, width and resolution of the network based on a set of coefficients.
Head: The output of the last block is fed into a pooling layer, which performs global average pooling across the spatial dimensions of the feature maps. This is followed by a fully connected layer and an activation function. The final layer of the network is a softmax layer.

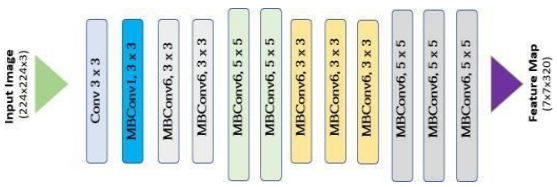

Figure 1. EfficientNet Architecture

In EfficientNetB7, the width of the network is

increased by a factor of 4 compared to EfficientNetB0, meaning that each convolutional layer has 4 times as many filters as the corresponding layer in EfficientNetB0.
For the HAR model, the final output from the LSTM layer is passed through a softmax activation function to classify the human action. In the final phases of our model architecture, we have used activation function "GeLU – Gaussian Error Linear Unit". It is a smooth approximation of the ReLU. It provides better performance than other activation functions like the ReLU, particularly for deeper neural networks. It is non-linear and has a range of values between 0 and infinity. The main advantage of the Gelu activation function over other activation functions like the sigmoid and hyperbolic tangent (tanh) functions is that it has been shown to perform better in deep neural networks.

The settings used to train models are displayed in Table 2.

Table. 2 Data Preprocessing & Augmentation Settings

Settings	Value
epoch	50
Batch size	128
Learning Rate	0. 0001
Drop Rate	0.5
Optimizer	RMSProp
Loss Function	Categorical crossentropy

Our proposed EfficientNetB7 model achieved a higher accuracy of 83.5% which is higher than all the other models discussed above. The HAR model trained on the custom action dataset achieved an accuracy of 81.2%.Table 3 shows the accuracy of different methods applied on the AffectNet dataset.

Table. 3 Accuracy Chart of Models Applied on AffectNet Dataset

Reference	Method	Accuracy
Kim et al., 2020	Xception	75.0%.
Gera et al., 2021	CCT	66.0%.
Zheng et al., 2022	POSTER	67.3%
Luna-Jiménez et al., 2021	STN	68.6%
Schoneveld et al., 2021	DeepFEVER	65.4%
Georgescu et al., 2019	CNN	63.3%
Zeng et al., 2022	Meta-Face2Exp	64.2%
Houshmand et al.,	VGG, ResNet	50.1%
2020,		49.2%
Ours	EfficientNetB7	83.5%
Ours	CNN-LSTM (HAR)	81.2%

V. Ablation Study

On the AffectNet dataset, an ablation study is conducted to demonstrate various optimizers performance. Table 4 shows the performance of Adam, SGD and RMSProp optimizer on the dataset. However, RMSProp optimizer showed improved results.

VI. Model Training and Evaluation

For model building and training, we use a python language. For training and evaluation Google Colab notebooks used Google's cloud server. For training, we also used Google's GPU and TPU where Google Colab is free to use. Transfer learning was applied using a pre-trained EfficientNetB7 model with ImageNet weights for FER, and a hybrid CNN-LSTM model for HAR was trained from scratch using the curated action dataset. To improve the overall performance of the model, the weight of a pre-trained model is being fine-tuned on the new dataset.

Table. 4 Accuracy Chart of Optimizers on AffectNet Dataset

Optimizer	Accuracy
Adam	82.4%
SGD	82.7%
RMSProp	83.5%

A popular optimization algorithm that is used in deep learning called Adam which helps to achieve the updated weights of neural networks during training. It includes adaptive learning rates and momentum. The adaptive learning rate in Adam allows the algorithm to automatically adjust the learning rate during training based on the gradient magnitudes. The momentum in Adam allows the optimizer to keep track of the gradient history and adjust the weight updates accordingly. This helps the optimizer to avoid getting stuck in local minima and to converge more quickly to the global minimum. The test and validation accuracy plot on our dataset for Adam optimizer is displayed in Fig. 1. The accuracy achieved by applying Adam optimizer on our training model turned out to be 82.4%.

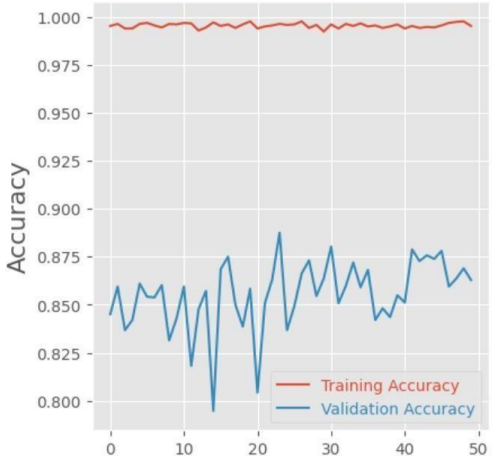

Figure 2. Training vs Validation Accuracy on Adam Optimizer

To minimize the loss function of the neural network during training, an SGD optimizer is used. It helps to compute the gradient of the loss function with the use of parameters for a small subset of the training data and then update the model parameters. The test and validation accuracy plot on our dataset for SGD optimizer is displayed in Fig. 2. The accuracy achieved by applying Adam optimizer on our training model turned out to be 82.7%.

Figure 3. Training vs Validation Accuracy on SGD Optimizer

RMSProp is a gradient descent optimization algorithm commonly used in neural networks to update the weights of the model during the training process. It works by maintaining a moving average of the squared gradient values for each weight. This allows the optimizer to scale the learning rate adaptively for each weight based on the magnitude of the gradients, which helps to prevent the optimizer from getting stuck in local optima. The test and validation accuracy plot on our dataset for RMSprop optimizer is displayed in Fig. 3. The accuracy achieved by applying RMSProp optimizer on our training model turned out to be 83.5%.

Precision, recall, F1 score, and support are commonly used metrics in classification tasks, which are used to evaluate the performance of a model in predicting the correct class labels. These metrics are often reported together to provide a more comprehensive evaluation of a classification model's performance.

Precision: Precision is the ratio of true positives (TP) to the total number of positive predictions (TP + false positives, FP). In other words, precision measures how many of the predicted positive examples are actually positive. A high precision means that the model is correctly identifying positive examples and not making many false positive predictions.

Recall: Recall is the ratio of true positives to the total number of actual positive examples (TP + false negatives, FN). Recall measures how many of the actual positive examples were correctly identified by the model. A high recall means that the model is correctly identifying positive examples and not missing many actual positive examples.

F1 Score: F1 I is the harmonic mean of precision and recall, where the formula is:

2(precision * recall) / (precision + recall)***.**

It provides a balance between precision and recall that is a good metric when uneven distribution of classes in data.

Support: Support is the number of actual occurrences of the class in the data. In other words, we can say that support is a number of examples that occur in a dataset and belong to the particular class.

VII. IMPLEMENTATION

The generated model can be exported as TensorFlow lite version for application purposes. We have used the Haar Cascade classifier for face detection. Haar Cascade Classifier is a machine learning-based object detection algorithm used to identify and locate objects in an image or video. It works by detecting features in an image using a set of classifiers, each of which corresponds to a specific feature or pattern in the image. The algorithm then applies these classifiers to sub-regions of the image at different scales and orientations, and if a match is found, the algorithm outputs the location of the object. This process is repeated until all the sub-regions of the image have been searched, and all possible objects have been detected. Our proposed solution gives the best interference time among other models. It is also significant that the performance of our proposed method is better than the other methods for the AffectNet dataset where their accuracy results vary between 52- 75%.

HAR Ablation Study Results:

For HAR, the following experiments were conducted: Testing with only CNN for spatial feature extraction resulted in reduced temporal understanding, achieving an accuracy of 75.3%.

Incorporating LSTM layers improved the temporal feature representation and led to better performance, achieving 81.2% accuracy.

Table. 5 Accuracy Chart of Optimizers on AffectNet Dataset

Optimizer	Accuracy
Adam	78.4%
SGD	79.7%
RMSProp	81.2%

The RMSProp optimizer achieved the highest accuracy due to its ability to adapt learning rates for each parameter dynamically, preventing overfitting and ensuring better convergence.

VIII. CONCLUSION

This paper presents a superior method for recognizing facial expressions using the

EfficientNetB7 model on the AffectNet dataset. The article discusses techniques for image preprocessing and data augmentation, as well as the architecture of the EfficientNetB7 model. The model achieves an accuracy of 83.5%, surpassing other models used on the AffectNet dataset. The EfficientNetB7 model achieved an accuracy of 83.5% for FER tasks, while the CNN-LSTM model demonstrated an accuracy of 81.2% for HAR tasks, both outperforming existing methodologies. The integration of FER and HAR enables seamless real-time mapping of facial expressions and human actions to virtual characters, making the system suitable for interactive VR applications. Our solution has been implemented in an application that can detect facial expressions in real time and map them to virtual characters' faces and action mapping as well. Our future work aims to further improve the model's performance.

REFERENCES

[1] Fard, A. P., Hosseini, M. M., Sweeny, T. D., & Mahoor, M. H. (2024). AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels. arXiv preprint arXiv:2410.22506.

[2] Gursesli, M. C., Lombardi, S., Duradoni, M., Bocchi, L., Guazzini, A., & Lanata, A. (2024). Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets. IEEE Access.

[3] Waldner, D., & Mitra, S. (2024). Pairwise Discernment of AffectNet Expressions with ArcFace. *arXiv preprint arXiv:2412.01860*.

[4] Feng, B., & Zhang, H. (2024, October). Expression Recognition Based on Visual Transformers with Novel Attentional Fusion. In *Journal of Physics: Conference Series* (Vol. 2868, No. 1, p. 012036). IOP Publishing.

[5] Feng, B., & Zhang, H. (2024, October). Expression Recognition Based on Visual Transformers with Novel Attentional Fusion. In *Journal of Physics: Conference Series* (Vol. 2868, No. 1, p. 012036). IOP Publishing.

[6] Patra, S., Kuanar, S. K., & Punuri, S. B. (2024, July). MHAN-FERW: Multi-stage Hierarchical Attention Network for Facial Emotion Recognition in Wild. In *2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)* (pp. 1-6). IEEE.

[7] Kim, J. H., Poulose, A., Reddy, C. S., & Han, D. S. (2021, June). Segregation of AffectNet Dataset for Facial Emotion Recognition. In Proceedings of Symposium of the Korean Institute of communications and Information Sciences (KICS) (pp. 1210-1211).

[8] Gera, D., & Balasubramanian, S. (2021). Consensual collaborative training and knowledge distillation based facial expression recognition under noisy annotations. arXiv preprint arXiv:2107.04746.

[9] Zheng, C., Mendieta, M., & Chen, C. (2022). POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition. arXiv preprint arXiv: 2204.04083.

[10] Luna-Jiménez, C., Cristóbal-Martín, J., Kleinlein, R., Gil-Martín, M., Moya, J. M., & Fernández-Martínez, F. (2021). Guided spatial transformers for facial expression recognition. Applied Sciences, 11(16), 7217.

[11] Schoneveld, L., & Othmani, A. (2021, September). Towards a general deep feature extractor for facial expression recognition. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 2339-2342). IEEE.

[12] Georgescu, M. I., Ionescu, R. T., & Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. IEEE Access, 7, 64827-64836.

[13] Zeng, D., Lin, Z., Yan, X., Lio, Y., Wang, F., & Tang, B. (2022). Face2exp: Combating data biases for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20291-20300).

[14] Houshmand, B., & Khan, N. M. (2020, September). Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM) (pp. 70-75). IEEE.

[15] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., & Chen, T. (2018). Recent advances in convolution neural networks, Pattern recognition, 77, 354-377.

[16] Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR

[17] Do, H. N., Trang, K., Vuong, B.Q., Tran, V. S., Mai, L., Vo, M., T., & Nguyen, M. H. (2020). Automatic facial expression recognition system using convolution neural networks. In 7th International Conference on the Development of Biomedical Engineering in Vietnam (BME7) Translational Health Science and Technology for Developing Countries 7 (pp. 473-476). Springer Singapore. Hyperparameter Tuning and Cross-Validation," Arab. J. Sci Eng., vol. 45, no. 12, pp. 10859-10873, 2020, doi: 10.1007/s13369-0204907-7