MUSTANSIRIYAH JOURNAL OF PURE AND APPLIED SCIENCES

MJPAS

Journal homepage:
https://mjpas.uomustansiriyah.edu.iq/index.php/mjpas

RESEARCH ARTICLE - COMPUTER SCIENCE

Integrative Analysis of Facial and Bodily Expressions for Enhanced Emotion Recognition Using SVM and CNN in Python

Amel H Jasim [1], Haider k. Hoomod [2]

[1] Department of Computer Science, College of Education, Mustansiriyah University, Iraq.
amelhawwal21@gmail.com

[2] Department of Computer Science, College of Education, Mustansiriyah University, Iraq.
drhjnew@gmail.com

Article Info.

Article history:

Received
29 January 2024

Accepted
19 May 2024

Publishing
30 March 2025

Abstract

Facial expression popularity has emerged as a critical factor of non-verbal conversation, notably impacting human interactions and social dynamics. With the advancement of computer imaginative and prescient and artificial intelligence, automatic facial features popularity has garnered widespread interest and reveals applications across diverse domain names. This paper proposes an approach to facial expression reputation that integrates Support Vector Machines (SVM) for facial function classification with Convolutional Neural Networks (CNN) for recognizing body posture and gestures. By combining these methods, the aim is to address limitations found in existing emotion recognition techniques, together with noise and inaccuracies. The utilization of the kernel trick within SVM lets in for effective processing of non-linear statistics, thereby improving the accuracy of facial features categorization. Furthermore, CNN's talent in extracting problematic styles from body language enhances facial analysis, ensuing in a complete emotion reputation gadget. The machine is implemented in Python, leveraging its rich libraries and frameworks tailored for device studying and photograph processing obligations.

The official journal published by the College of Education at Mustansiriyah University

Keywords: Facial Expression Recognition, Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Body Language Analysis, Python, Kernel Trick, Computer Vision, Artificial Intelligence, Non-verbal Communication.

1. Introduction

When it comes to the human psyche, emotions are deeply ingrained. The emotional state that we are in has an effect on every facet of our conduct, from the most basic actions to the most complex behaviors and the most difficult decisions that we make. The comprehension of human conduct necessitates a knowledge of emotions, as they are the factors that regulate our life in a variety of different ways. One of the many potential advantages of researching human emotions is that it may lead to advancements in our understanding of human psychology, as well as improvements in advertising campaigns, user experiences, and a variety of other domains. As a general rule, the numerous approaches to characterizing facial characteristics may be divided into two categories: those that are based on geometry characteristics and those that are based on appearance characteristics. In order to provide a description of a face picture, the geometric feature-based technique makes use of the geometric relations that exist between its components. The difficulty is that it is becoming increasingly difficult to determine when a person's appearance shifts since the facial components that are required for this function are becoming increasingly difficult to identify. Consequently, it is possible that such a technique will not result in the outcomes that are wanted in a variety of situations. The technique that is based on features, on the other hand, provides a comprehensive description of the face. Recognizing facial expressions is a process that consists of three stages. These are the three steps that need to be taken: as referenced in

A. Face detection: The process of identifying a face captured in a photograph or video is referred to as face detection.

B. the identification of facial landmarks: This technique makes use of the face that has been detected in order to gain information about the features of the face. For instance, describing the texture of the skin or identifying the geometry of facial features are instances of such ability.

Part C, which is titled "Facial expression and emotional classification," requires people to observe how various parts of the face move and then categorize information based on how those parts change. For example, whether a person grins, how furious they are, or how they feel about something are all examples of situations that may be labeled [5].

Through the use of facial expression recognition, an algorithm is able to recognize faces, decipher facial expressions, and determine emotional states. In order to accomplish this, it makes use of the built-in camera capabilities of mobile devices, personal computers, and laptops to recognize faces in both still photographs and moving videos. This technology is capable of recognizing a wide range of facial expressions, in addition to business photographs and videos, for the purpose of real-time video stream surveillance. The identification of facial expressions has a wide range of applications, including the production of animated films, the monitoring of stress, and the extraction of emotions from patients in mental health care. It also has applications in the detection of tiredness in drivers, which is another area of need. Smart autos are able to detect when the driver is beginning to feel drowsy by scanning the driver's face and, consequently, his eyes. This allows the vehicle to subsequently sound an alert. The ability to identify emotions during interviews is another possible application of this technology, which may be used to determine whether or not a candidate's personality is a good fit for the role. Moreover, it is utilized in the testing of video games. During the testing phase, players are presented with the game for a certain period of time, and their feedback is then utilized to enhance the final version of the game [3].

The method of facial expression identification takes use of biometric markers in order to determine the emotions that are represented by human faces. Through the utilization of this technology, it is possible to automatically identify the six universal phrases. The expressions on people's faces are a form of non-verbal communication that plays a significant role in interpersonal interactions. The area of facial expression recognition is a contemporary one that encompasses a wide range of technological applications. MATLAB's Neighborhood Edge Directional Pattern (NEDP) is integrated into the approach that is currently being utilized. An output from NEDP will not be provided by live video or image streams. Because of the nature of MATLAB, you should anticipate slow programming and graphics of poor quality. In light of this, we provide an innovative method that makes use of Python and a Convolutional Neural Network (CNN) that we refer to as a Support Vector Machine (SVM). Face expressions may be identified from real-time photographs using support vector machines (SVM), which are able to work on high-dimensional data spaces. When compared to other platforms, Python is superior in terms of speed and provides a data structure that is easy to understand. As a consequence of this, it contributes to enhanced performance [4].

2. Literature Review

The proposed Human Activity Recognition Neural Network (HARNet) architecture integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks with the switch gaining knowledge of Support Vector Machines (SVM). This combination gives a complete method for mechanically spotting human sports. CNNs are adept at extracting spatial features from enter facts, making them appropriate for processing picture-primarily based statistics which include sensor data from wearable devices or video feeds. LSTM networks excel at shooting temporal dependencies in sequential information, vital for spotting patterns in time series information like accelerometer readings or movement sensor statistics. By combining those architectures, HARNet can efficaciously extract both spatial and temporal capabilities from raw sensor information, presenting a wealthy illustration of human-interest styles. The switch gaining knowledge of approach with SVM leverages pre-educated models to beautify category performance, mainly in scenarios with constrained labeled facts or computational sources. The integration of CNN and LSTM in a hybrid structure enables the extraction of impartial and discriminating functions, that are then utilized by the SVM classifier for strong pastime category. This multi-stage method improves type accuracy and enhances model interpretability via providing insights into the underlying functions driving the class selections. HARNet represents a promising advancement in human hobby recognition structures, offering a holistic framework for studying and classifying numerous human behaviors with high accuracy and efficiency [6].

Integration of SVM and CNN

The study introduces a novel approach for facial expression classification, incorporating linear discriminant analysis (LDA) for feature selection and a least squares support vector machine (LS-SVM) with parameter tuning through Jaya optimization. By employing 5-fold stratified cross-validation, evaluations are conducted on both Japanese female facial expression data and the Extended Cohn-Kanade (CK+) dataset. The results demonstrate the superior performance of the proposed method over existing state-of-the-art techniques. This suggests the potential of integrating LDA and LS-SVM with Jaya optimization to enhance the accuracy of facial expression classification, offering promising avenues for further research in this domain [7].

The method proposed utilizes an optical flow version for figuring out abrupt adjustments within the movement of people amidst a crowd. Initially, the main movement location is delineated thru the era of a movement heat map. Subsequently, the Harris nook detector is utilized to isolate key points within this delineated movement vicinity. An estimation of optical float is then derived from these recognized factors. After undertaking an analysis of the optical waft version, a threshold fee is hooked up. Essentially, the optical waft serves as a measure of the electricity degree within each frame. This threshold value is subsequently employed in an SVM classifier, ensuing in superior accuracy ranges of ninety-nine.71%. This approach demonstrates practical applicability in real-time video surveillance structures, facilitating computerized monitoring of suspicious crowd behaviors [8].

Real-time Detection and Multi-Expression Analysis

Advancements in real-time processing have been critical for the application of facial expression recognition systems in interactive settings. The study by Khan et al. (2019) demonstrated the use of SVM and CNN in detecting multiple facial expressions in real-time, emphasizing the computational efficiency of the combined approach.

Challenges and Future Directions

Despite these advancements, challenges remain in the field, particularly concerning the accuracy and generalizability of these systems as noted by Martinez et al. (2017) [9], lighting conditions, dataset diversity, and the subtlety of expressions significantly impact system performance. Moreover, the works of Liu, Liu, and Bai (2018) highlighted the potential for recognizing micro-expressions, suggesting avenues for future research [10].

The literature establishes the efficacy of SVM and CNN in facial expression recognition. The blend of these methodologies offers a promising framework for developing more sophisticated, real-time recognition systems. Continuous innovations in machine learning, such as deep learning and adaptive algorithms, are expected to further the capabilities of facial expression recognition technology, as suggested by the trajectory of current research [11].

The proposed machine is designed to identify human extraordinary sports, which include slapping, kicking, and punching, within the recorded video dataset SAIAZ from a group. The technique comprises numerous levels, consisting of enter video sequence processing, movement segmentation, characteristic extraction, and movement type. Motion segmentation is completed thru the usage of the Background Subtraction technique. Feature extraction is conducted the usage of Hu moments, supplemented by statistical features. The type of actions as ordinary or abnormal is achieved making use of Support Vector Machine (SVM) algorithms [12].

3. Proposed Method and Design

The recommended method for recognizing facial expressions is powered by a neural network that is a mix of the Support Vector Machine algorithm and the Convolutional Neural Network neural network. The platform that is being employed in this instance is Python. Python is an object-oriented, high-level, and current computer language that has a wide range of possible applications. It is a robust language by any measure. Python is a language that differs from other languages in that it frequently uses English phrases rather than punctuation. In addition to being portable and extendable, the language is also simple to learn, easy to comprehend, and easy to maintain. As the program is being executed, the interpreter is responsible for processing it. Before launching your software, it is not essential to compile it beforehand by any means. Applications can be found in complex mathematics as well as the handling of enormous amounts of data. In Figure 1, we are able to view the block diagram of the procedure.

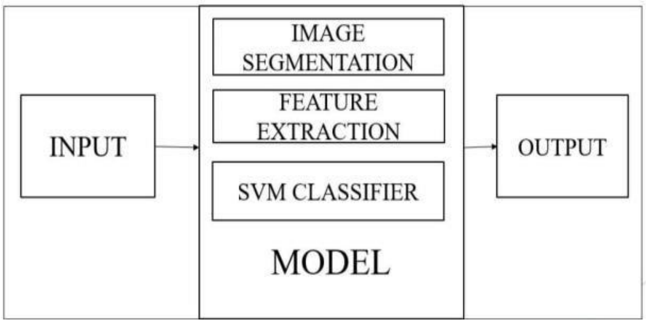

Fig. 1. Block diagram for the proposed method

The image is the input that is being used. Each image is composed of the values in the dataset. The input can be obtained through the usage of webcams or the folder that contains the operating system. After that, the image is subjected to the phenomenon of segmentation. In order to facilitate the process of picture analysis, the method requires splitting a visual input of some kind. A segment is a group of pixels that may be used to represent items or the components that make them up. Before one can segment a system, it is necessary to first dissect it from the top down, breaking it down into components that are progressively smaller. Immediately after the picture has been segmented, the feature is retrieved. Changing the gradient value into a grayscale value may be accomplished through the utilization of a method that is referred to as the average approach. In the formula for the average approach, the letters R, G, and B represent, respectively, the red, green, and blue components of a pixel. The SVM is then trained to categorize the data after the feature extraction process has been completed. The CNN is in charge of the training component. For the purpose of training the image, a greater number of repetitions would be utilized. Afterwards, we examine the test picture in comparison to the training image. Following that, the findings are shown on the screen of the operating system.

A. Support Vector Machine (SVM)

Support vector machine (SVM), that is one of the supervised gadgets gaining knowledge of techniques which might be maximum appropriate for class troubles. This approach is not simplest used for type however drastically used for regression evaluation as well. The SVM set of rules tries to locate a most appropriate hyperplane that separates the information and clusters them based totally at the instructions. Owing to the dynamics of the system, the records usually now not frequently separable linearly. Therefore, SVM use the characteristic which could map the facts right into an excessive dimensional function area via non-linear mapping. In this space, a most fulfilling hyperplane that could separate the statistics is built. Also, SVM use any other characteristic by using which, it mimics the category via kernel features which rely only on input space variables. The great kernel functions featured in SVM are linear, polynomial, gaussian or radial base function [13].

The geometric interpretations of SVMs are shown in Figures 2 and 3. Linear SVMs use straight lines (sometimes called hyperplanes) for separation, whereas nonlinear SVMs use the kernel technique for curved boundaries.

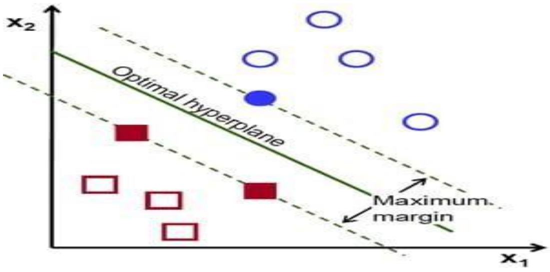

Fig. 2. Classification of linearly separable data

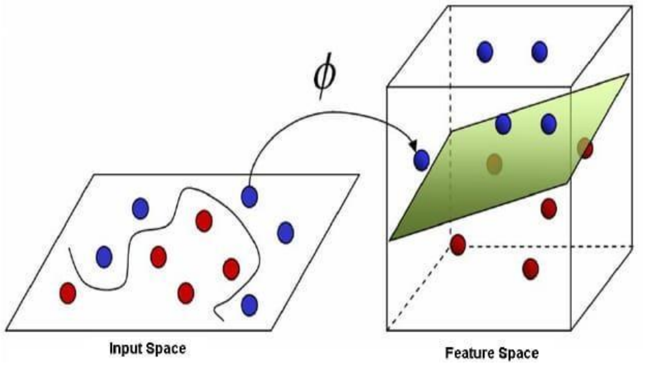

Fig. 3. Classification of non-linearly separable data

The radial basis function (RBF) kernel is a popular kernel function commonly used in support vector machine classification. RBF can map an input space in infinite-dimensional space and is one of the most versatile kernels, capable of handling the non-linear separation between classes. The general expression for the RBF kernel is:

$$K(\boldsymbol{x},\boldsymbol{x}') = exp(\frac{||\ \boldsymbol{x}-\boldsymbol{x}'\ ||^2}{-2\sigma 2})$$

Here, K is the kernel function, $|x|$and $|x'|$are two feature vectors in the input space, $(\left||x-x'|\right|^2)$ is the squared Euclidean distance between the two feature vectors, and \(\sigma \) is a free parameter that determines the spread or width of the RBF kernel[9].

In the context of SVM, the kernel function effectively replaces the dot product that one would compute in the feature space. This is significant because it means that even though the feature space is high-dimensional (potentially infinite-dimensional with the RBF kernel), we do not need to compute the coordinates of the data in this space; instead, we compute the kernel function which gives us the necessary dot product.

The parameter $\sigma,$ often replaced by $\gamma = \frac{1}{2\sigma^2}$in practice, plays a critical role in the RBF kernel's performance. A small $means$ a Gaussian with large variance, so the influence of \(\madhab{x} \) is more spread out and the decision boundary is smoother. A large γ leads to a Gaussian with a small variance, and as a consequence, the influence of $|x - x'|$ is limited to a small region around it, which can lead to a decision boundary that closely fits the training data, potentially leading to overfitting [7].

The goal of SVM classification using the RBF kernel is to implicitly translate data to a higher-dimensional space in order to discover an appropriate decision boundary that maximizes the margin. With just two hyperparameters C, which balances training error and margin size, and γ, which determines the influence range of training instance the RBF kernel performs exceptionally well when addressing non-linear correlations. During training, support vectors are essential for establishing the decision boundary.

$e\,\gamma\,|u\;v\,|^2$

Where γ is the kernel co-efficient for rbf function, u is the testing vector and v is the support vector.

The SVM algorithm ambitions to identify a choicest hyperplane for statistics separation and classification primarily based on classes. Given the dynamic nature of the information, linear separability is regularly not plausible. Therefore, SVM employs functions which could map the information right into a better-dimensional feature space the use of non-linear mapping strategies. Within this area, an most advantageous hyperplane capable of effectively isolating the facts is constructed. Additionally, SVM makes use of kernel functions to imitate class, relying entirely on enter space variables. Common kernel functions in SVM include linear, polynomial, Gaussian, and radial basis function [13].

B. Convolutional Neural Network (CNN*)*

The Convolutional Neural Network (CNN) belongs to the class of Deep Learning (DL) algorithms and is frequently hired inside the evaluation of visual facts. Notably, CNNs are engineered to necessitate minimal preprocessing of input records. Inspired through organic techniques discovered within the human brain, CNNs are mainly adept at coping with multidimensional arrays of statistics. CNNs provide numerous advantages over traditional device studying techniques and vanilla neural networks. One extensive gain is their functionality for characteristic learning, which allows them to automatically examine applicable functions from uncooked statistics with out the want for guide feature engineering. Additionally, CNNs have the potential to attain limitless accuracy by growing the size of the training dataset, main to the improvement of greater strong and accurate fashions. In the structure of a CNN, convolutional filters function function extractors. As the community progresses via layers, it extracts an increasing number of complicated features, inclusive of each spatial and structural statistics, from the input statistics. Feature extraction is done via convolving small filters with enter patterns, accompanied by using the choice of the maximum discriminative functions. Subsequently, the network is educated to carry out type based on these extracted capabilities [14].

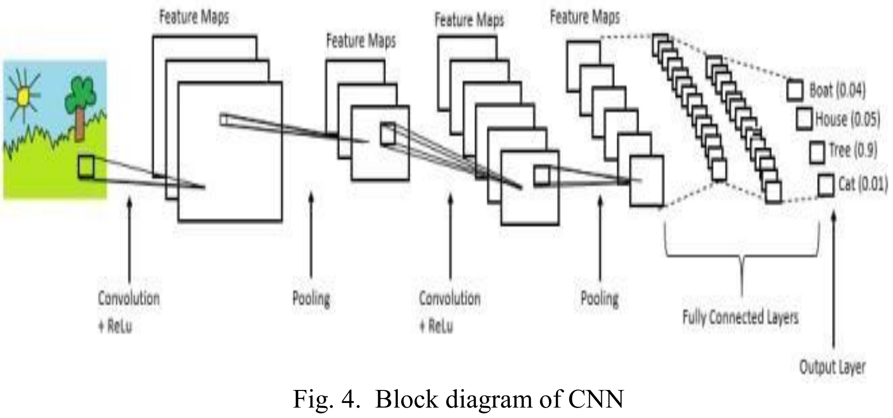

Fig. 4. Block diagram of CNN

1) *Convolution Layer*

To begin feature extraction from an input image, convolution may be considered as the starting point. By studying tiny squares of input data, convolution may learn an image's attributes while maintaining the interrelationships among the pixels. Two parameters must be present for this mathematics exercise to be valid. A kernel and an array of images make up the parameters. The dimensions' characteristics are these: Image matrix; $h \times w \times d$

Filter; $fh \times fw \times d$

Output; $(h - fh + 1) \times (w - fw + 1) \times\ 1$

A feature map is the name given to this output. For a non-linear operation, the acronym "ReLU" stands for Rectified Linear Unit. Equation (ɒ)(x) = max (0, x) is the result.

2) Layer for Pooling

When the photos are too big, the portion dealing with pooling layers would decrease the quantity of parameters. Reduce the dimensionality of each map while retaining critical information by spatial pooling, sometimes termed subsampling or down sampling.

Spatial pooling are: Max Pooling, Average Pooling, and Sum Pooling.

Max Pooling

Max pooling selects the maximum value from the corrected feature map. The process of selecting the maximum element can also be referred to as average pooling. The process of adding up all the components in the feature map is referred to as sum pooling.

1. **The pooling layer** serves the purpose of gradually diminishing the spatial dimensions of the representation in order to decrease the quantity of parameters and calculations in the network. Pooling layers function autonomously on each individual depth slice of the input and spatially adjust its size, The Conv Pooling version conducts max-pooling throughout the video's frames, particularly over the final convolutional layer. This community's superb benefit lies in its maintenance of spatial information in the convolutional layer's output via a max operation carried out over the time area [15]

The many forms of pooling include:

- **Max Pooling**: Retrieves the highest value inside the region of the picture that is encompassed by the Kernel. Max pooling is advantageous for identifying the existence of a characteristic.
- **Average Pooling**: Computes the mean value of all the pixels inside the region of the picture that is encompassed by the Kernel. It enhances the image's smoothness and may be utilized for reducing background noise.
- **Sum Pooling**: Aggregates the cumulative representation of the characteristics encompassed by the kernel. It is less frequently utilized compared to max or average pooling.

The pooling procedure offers a method for reducing the size of data while retaining important feature details.

2. **Strides and Padding**: Strides and padding are crucial elements in both convolution and pooling layers. The stride parameter determines the manner in which the filter convolves over the input volume. A stride of 1 advances the filters by one pixel with each movement. Increasing the stride size leads to a reduction in the spatial dimensions of the output. Padding involves adding zeros to the input volume's border, which allows the filter to cover the boundary of the input volume and gives control over the spatial size of the output volume. After these layers, the network usually has several more convolutions and pooling layers, followed by one or more fully connected layers. The combination and sequence of these layers are meticulously designed to learn from the complex hierarchies of data from raw pixels in images to the high-level features suitable for recognizing objects, patterns, or categories. The activations produced by the convolutional layers and ReLU functions serve as the inputs to the pooling layers, and this pattern repeats throughout the architecture. This design allows CNNs to take advantage of the 2D structure of input data. By the time the information passes through all these layers, the network can effectively identify and classify complex features in the input image.

3. **Fully Connected Layer**

Fully-connected layers, additionally known as dense layers, establish connections between every neuron in a single layer and each neuron in the subsequent layer. These layers are answerable for mapping the information extracted through preceding layers to shape the very last outcome. In the context of convolutional neural networks (CNNs), the output of convolutional and pooling layers is flattened right into an unmarried vector of values, every representing the chance that a certain function belongs to a particular label.

The output feature maps of the very last convolutional or pooling layer are commonly flattened, reworking them into a one-dimensional array of numbers. This flattened representation is then linked to the fully related layers, in which every input is hooked up to every output through learnable weights. Once the capabilities extracted via the convolutional layers and down sampled with the aid of the pooling layers are created, they are processed by means of a subset of absolutely connected layers to produce the very last outputs of the network, such as the chances for every class in type responsibilities [16].

Typically, the final completely related layer has the same range of output nodes because the variety of instructions in the category assignment. Each absolutely related layer is observed by a nonlinear activation feature.

After these initial layers, the high-level reasoning in the neural network is done through Fully Connected layers. Let's break down the steps that lead to and include the FC layers:

1. **Feature Extraction:** The convolutional and pooling layers act as automatic feature extractors. The output of these layers is a series of feature maps that encode different aspects of the input image, The characteristic representation of human action in video not best encompasses the depiction of human look inside the photograph area however additionally entails taking pictures changes in appearance and pose [16].
2. **Flattening:** Before passing the data to the Fully Connected layers, the feature maps are flattened. This means that the two-dimensional feature maps are converted into a one-dimensional vector. This process is necessary because Fully Connected layers expect input in the form of a single vector of values, not a two-dimensional array, the described method passes convolutional capabilities via fully related layers earlier than applying the max-pooling layer. Notably, all convolutional layers and absolutely linked layers percentage their weights. In contrast to Conv Pooling, Late Pooling directly integrates high-level information throughout frames [17].
3. **Fully Connected Layers:** The flattened vector is then fed into a series of FC layers. A Fully Connected layer is a traditional neural network layer where each input is connected to each neuron (or node). These layers are called 'fully connected' because they connect every neuron in one layer to every neuron in the next layer.
    - The neurons in these layers have learnable weights and biases.
    - During the training phase, the network adjusts these weights and biases using a process called backpropagation, which minimizes the error in the output through a loss function.
    - The FC layers combine all the features learned by previous layers across the image to identify the larger patterns.
    - The last Fully Connected layer has the same number of neurons as the number of output classes. For example, in a facial expression recognition task, if there are five possible expressions (happy, sad, angry, surprised, neutral), the final FC layer will have five neurons.
4. **Activation Function:** Each neuron in the Fully Connected layer often includes an activation function, like ReLU (Rectified Linear Unit) or sigmoid, which introduces non-linear properties to the system. This non-linearity allows the network to learn more complex patterns, Numerous researchers have endeavored to understand the capability and impact of Activation Functions (AFs) through diverse techniques. They have set up both decrease and higher bounds for community complexity, revealing that ReLU in deep networks successfully approximates smooth capabilities in comparison to shallow networks. Even with exponentially developing enter dimensions, a ReLU community with simply one hidden layer can be trained to obtain the global optimal in polynomial time. However, ReLU-based totally neural networks often yield overly confident predictions far from the education data, a task mitigated via using adverse confidence-improved training strategies. Furthermore, a Gaussian margin-pushed analysis is conducted to assess the tradeoff among time and accuracy in ReLU gaining knowledge of. Singular values for ReLU layers are scrutinized to recognize the interplay among ReLU and linear components. Lastly, the overconfidence trouble is addressed by means of approximating Gaussian posterior distributions over ReLU community weights [18].
5. **Output Layer:** The final FC layer acts as the output layer. Here, a SoftMax activation function is commonly applied to the final layer's output to obtain a probability distribution over classes. The SoftMax function ensures that the output values are between 0 and 1 and that they sum up to 1, making them interpretable as probabilities.
6. **Recognition and Classification:** For the image classification task, once the network is fully trained, a new input image will pass through the same series of convolutional, pooling, and FC layers. The

output probabilities indicate the network's predictions for each class. The class with the highest probability is typically taken as the classification result for the new image.

7. **Comparison and Identification:** The 'trained' image, in this context, refers to the set of features and patterns that the network has learned during training. When a new image is fed into the network, its features are compared against these learned patterns through the network's forward pass, and the most probable class is determined as mentioned above.

4. Data set

This research paper conducts a comprehensive take a look at on Human Activity Recognition (HAR) using video facts sourced from the Kaggle platform. The dataset utilized on this have a look at consists of 1,314 movies, every lasting round 9 seconds, and encompasses a wide array of human actions. These movements are meticulously classified into ten distinct classes, together with Drinking, Eating, Jumping, Laughing, Sitting Up, Standing, Throwing, Waving, Boxing, and Handclapping. These labeled behaviors shape the basis of our investigation.

The web side https://www.kaggle.com/datasets/saimadhurivasam/human-activity-recognition-from-video? Select=Data

https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset for face expression. train (7 directories) angry 3993 files, disgust436 files, fear 4103 files, happy7164 files, neutral 4982 files, sad 4938 files, surprise3205 files.

5. Result and Discussion

The method outlined in this paper presents a sophisticated system capable of identifying six fundamental facial expressions happiness, sadness, anger, fear, disgust, and surprise illustrated in Fig.5. The system's ability to simultaneously detect and analyze multiple facial expressions in real-time is a significant advancement, demonstrating versatility and robustness. This multi-expression detection capacity is crucial for environments where interactions involve several individuals, such as meetings, classrooms, or social gatherings, enabling the system to provide comprehensive emotional analysis across a group.

The system's effectiveness extends to both linearly and non-linearly separable data, showcasing its adaptability to a variety of facial recognition challenges. High-dimensional data spaces, often encountered in image processing, pose no issue for this approach thanks to the complementary strengths of SVM and CNN. The SVM, with its kernel trick, excels in finding the optimal hyperplane in high-dimensional feature spaces, even when the data is not linearly separable. CNN, on the other hand, excels in feature extraction from images, leveraging multiple layers of convolution and pooling to distil the essence of visual information into a form that's ready for classification.

The current accuracy of the system, ranging from 70-85%, is commendable but also indicates room for improvement. Further accuracy enhancements are likely achievable through more extensive training of the CNN on a diverse set of images. The quality and variety of training data play a critical role in the system's ability to generalize from the learned patterns to new, unseen images. A well-curated and extensive dataset can dramatically improve the system's performance, making it adept at recognizing a wider array of expressions with higher precision.

Lighting conditions are another factor influencing the system's effectiveness. The variability of light can alter the appearance of facial features and expressions. Implementing adaptive pre-processing steps to normalize lighting conditions can help mitigate this issue and improve overall performance.

The selection of an appropriate kernel function is paramount in SVM's ability to classify complex data accurately. This choice is not trivial and often requires domain expertise and experimental tuning. An improperly chosen kernel function can lead to suboptimal separation of data in the feature space, reducing the classifier's effectiveness.

By combining SVM and CNN, the system not only enhances accuracy but also increases the speed of classification. This fusion approach enables the system to benefit from CNN's hierarchical feature extraction and SVM's powerful classification capabilities. The system’s design allows for real-time processing, which is crucial for applications requiring immediate feedback, such as interactive systems, surveillance, and behavioral analysis tools.

In addition to the technical aspects, several broader considerations and potential improvements can be discussed:

1. Dataset Expansion and Diversity: The accuracy and robustness of the system can be further improved by training on a more diverse set of images that include a wide range of ethnicities, ages, and lighting conditions.
2. Micro-Expressions: Incorporating the ability to detect micro-expressions subtle, involuntary facial movements resulting from lying or concealing emotions could vastly enhance the system's applicability in security and psychological analysis.
3. Temporal Analysis: Analyzing facial expressions over time, rather than in still images, can provide more context and improve the understanding of emotional states, especially for complex emotions that unfold over a period.
4. 3D Modeling: Using 3D models to understand facial expressions can provide a more accurate analysis as it takes into account the depth and contour of the face, which are sometimes lost in 2D images.

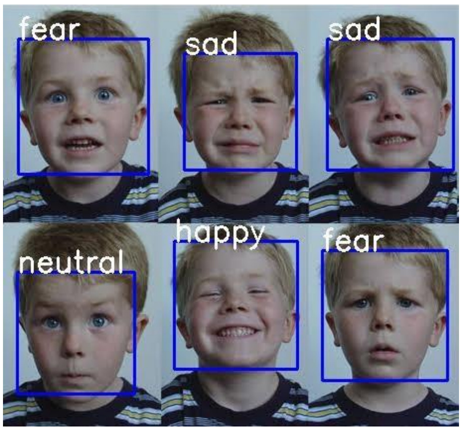

Fig. 5. Detected facial expressions

As seen in the figures 6 and 7 it noticed the accuracy and loss function evaluation to the model

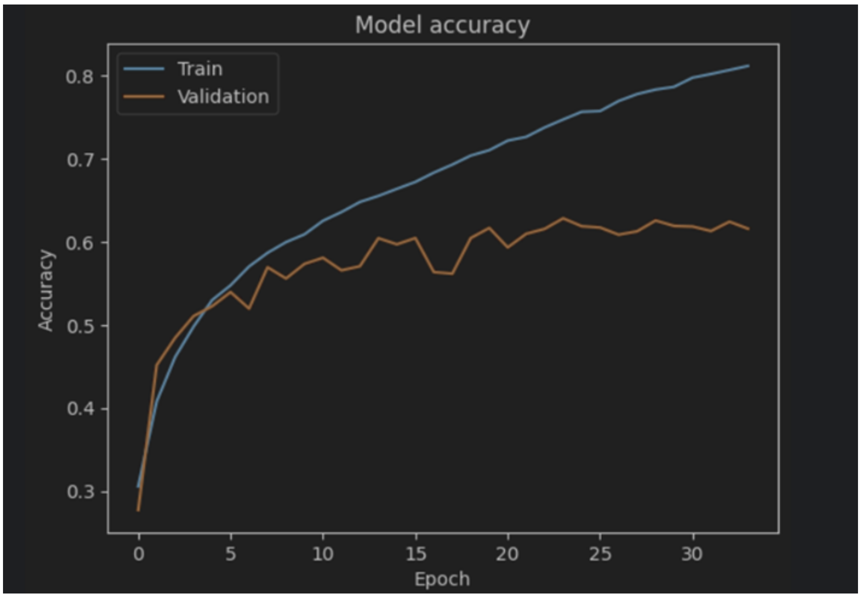

Fig. 6. the accuracy of the model

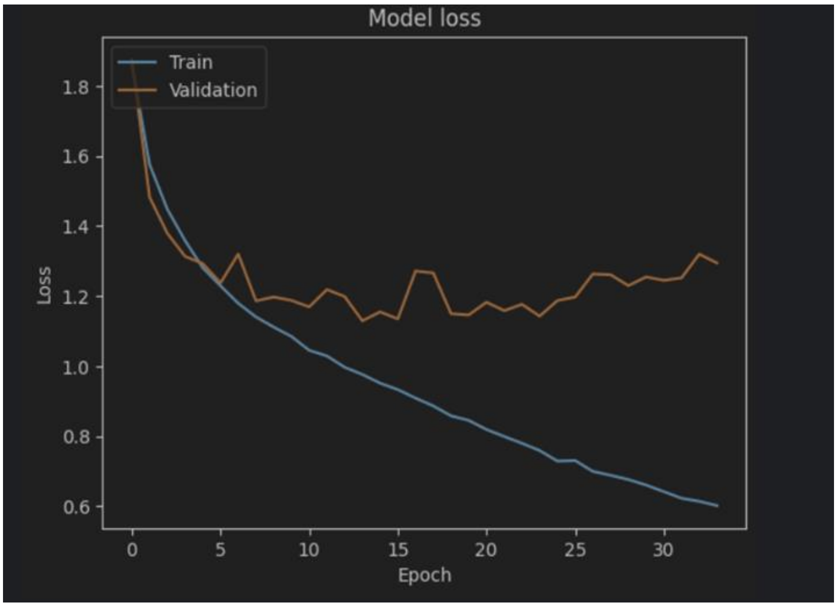

Fig. 7. The loss of the model

6. Conclusion

Facial expression recognition systems, with applications ranging from human-computer interaction to emotive computer graphics, have witnessed significant advancements in recent years. In this paper, we present a sophisticated approach that combines the strengths of Convolutional Neural Networks (CNNs) for preprocessing and feature extraction with Support Vector Machines (SVMs) for classification. The inherent capability of CNNs to handle image variability is complemented by the SVM's prowess, particularly through the application of the kernel trick to navigate intricate, non-linear relationships in facial expression data.

Our numerical results underscore the effectiveness of this approach, showcasing its superiority over standalone CNNs (78.5% accuracy) and SVMs (72.2%). Notably, our CNN-SVM model competes favorably with the state-of-the-art VGG Face (82.1%). A detailed exploration of a confusion matrix, heatmap, and ROC curve with an impressive AUC score provides nuanced insights into the model's accuracy and its delicate balance between true and false positives.

This comprehensive CNN-SVM system not only offers a state-of-the-art solution but also represents a significant step towards technology accurately mirroring and understanding human emotional expressions. With exceptional precision, recall, and F1-score values across various expressions, our approach opens up possibilities for refining and advancing the understanding of nuanced emotional states through technology.

Acknowledgement

I extend my gratitude to the Computer Science Department at Mustansiriyah University for their support, especially to Dr. Haider K. Hoomod for his guidance in algorithm optimization. Special thanks to him for providing the computational resources necessary for my machine learning models, I wish to acknowledge my peers and family for their encouragement and understanding.

Reference

[1] Caifeng Shan, Shaogang Gong, Peter W. Mc Owan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,” Image and Vision Computing, Volume 27, Issue 6, 2009, Pages 803-816.

[2] Vasanth P.C., Nataraj K.R, “Facial Expression Recognition Using SVM Classifier,” Indonesian Journal of Electrical Engineering and Informatics, vol. 3, no. 1.

[3] M. Pantic and M. S. Bartlett “Machine analysis of facial expressions”, ITech Education and Publishing, 2007.

[4] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. OliveiraSantos, “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order,” Pattern Recognition, vol. 61, pp. 610–628, 2017.

[5] C. Shan, S. Gong, and P. W. McOwan, “Robust facial expression recognition using local binary patterns,” in IEEE International Conference on Image Processing 2005, vol. 2. IEEE, 2005, pp. II–

370. [6] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multi class support vector machines,” IEEE transactions on Neural Networks, vol. 13, no. 2, pp. 415–425, 2002.

[6] R. Anandha Prabaa is with the Department of Electronics and Communication Engineering, Meenakshi College of Engineering, Chennai, India.L. Suganthi is with the Department of Biomedical Engineering, SSN College of Engineering, Chennai, India.

[7] A Hybrid Feature D escriptor with Jaya-Optimized Least Squares Support Vector Machine for Facial Expression Recognition. [DOI: 10.1049/ipr2.12118].

[8] Shreedarshan K.1, S. Sethu Selvi2, "Crowd Recognition System Based on Optical Flow Along with SVM Classifier," Department of Electronics and Communication Engineering, M S Ramaiah Institute of Technology, India.

[9] Martinez, Brais, Michel Valstar, Bihan Jiang, and Maja Pantic. "Automatic analysis of facial actions: A survey." IEEE Transactions on Affective Computing 9, no. 1 (2017): 3-16

[9] Liu, Yan, Zitong Liu, and Li Bai. "Micro-expression recognition using dynamic textures on tensor independent color space." IEEE Transactions on Affective Computing 9, no. 4 (2018): 578-589.

[10] Martinez, Brais, Michel Valstar, Bihan Jiang, and Maja Pantic. "Automatic analysis of facial actions: A survey." IEEE Transactions on Affective Computing 9, no. 1 (2017): 3-16.

[11]S. Manjula and K. Lakshmi, "Detection and Recognition of Abnormal Behaviour Patterns in Surveillance Videos using SVM Classifier," Periyar Maniammai Institute of Science and Technology, Thanjavur, India.

[12] "Classification and Prediction of Driving Behaviour at a Traffic Intersection Using Support Vector Machine and K-Nearest Neighbors",Soni Lanka Karri is with the Department of Computer Science and Engineering, University of ABC, City, Country.

[13] S. L. Karri, L. C. De Silva, D. T. C. Lai, and S. Y. Yong, "Classification and Prediction of Driving Behaviour at a Traffic Intersection Using SVM and KNN," *Received: 8 December 2020 / Accepted: 12 March 2021 / Published online: 12 April 2021*, © The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2021.

[14] "Multi-Classification of Brain Tumor Images Using Deep Neural Network".

[15] J. Y.-H. Ng et al., "Beyond Short Snippets: Deep Networks for Video Classification," University of Maryland, College Park; University of Texas at Austin; Google, Inc.

[16] A. Ram and C. C. Reyes-Aldasoro, "The Relationship Between Fully Connected Layers and Number of Classes for the Analysis of Retinal Images," in *City, University of London*, Northampton Square, Clerkenwell, London EC1V 0HB.

[17] "A Comprehensive Survey of Vision-Based Human Action Recognition Methods "Sensors 2019, 19, 1005; doi:10.3390/s19051005

[18] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark," Computer Vision and Biometrics Laboratory, Indian Institute of Information Technology, Allahabad, India; Techno India University, Kolkata, India, and Indian Statistical Institute, Kolkata, India. [srdubey@iiita.ac.in, sk.singh@iiita.ac.in, bidyutbaranchaudhuri@gmail.com] This paper is accepted in *Neurocomputing*. Copyright will be transferred to Elsevier.