Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/135/2024.18696

A Comprehensive Evaluation of Emotion Recognition Techniques: Model and Data Analysis

Guoguo Lyu[1,a,*]

[1]*College of Art & Science, University of Toronto, Toronto, Canada*
a. kelly.lyu@mail.utoronto.ca
**corresponding author*

Abstract: This study evaluates three prominent techniques for emotion recognition: Convolutional Neural Networks (CNNs), the Emotions in Context (EMOTIC) dataset, and feature extraction methods like Principal Component Analysis (PCA) and Local Phase Quantization (LPQ). By analyzing these methods across datasets such as Fuyeor Language (FER) 2013 and EMOTIC, the research highlights CNNs' ability to classify emotions from Electroencephalogram (EEG) signals with high accuracy, enhanced by PCA's dimensionality reduction. The EMOTIC dataset's fine-grained emotional categories, combined with contextual data, improved emotion detection in real-world settings, particularly when facial expressions alone were insufficient. LPQ further enhanced texture analysis in challenging environments with variable lighting. While CNNs demonstrated strong performance, challenges like real-time processing and generalization across diverse datasets remain. Future work should focus on integrating audio and physiological data and incorporating temporal information to detect dynamic emotional changes. These developments will help only in creating a better and generalized emotion detection system for the respective areas like healthcare, virtual reality, and interaction with the virtual world.

Keywords: Emotion Recognition, Convolutional Neural Networks (CNNs), Principal Component Analysis (PCA), EMOTIC Dataset.

1. Introduction

The release of the movie ‘Inside Out’ majorly brought to the limelight a significant aspect of life hitherto unnoticed which is the importance of emotions in the day-to-day activities of individuals and in all the interactions that people engage in [1]. The capability to comprehend other people’s emotions is the key to successful interaction. Recently, there has been a huge emphasis put on emotion recognition technology. This technology comprises of the reading of frowns, pitch of voices, posture, and other signs and symptoms related to human feelings. It is becoming more imperative to offer a customized and flexible Information and Communications Technology (ICT) service delivery. According to Ekman, it is crucial to decide on an individual’s behavior based on the ability to feel the emotions of the other person, and know how to recognize them or not [2]. For affective computing to be successful, three key processes are necessary: The three types of functions in emotional design are the identification of emotion, synthesizing, adapting and expressing emotions, and lastly, evoking emotions.

DOI: 10.54254/2754-1169/135/2024.18696

Recent developments in the areas of Emotion Recognition technologies have reached impressive evolution and improvement in the analysis of emotions, which are detected through different physical and behavioral response indicators. Starting with more basic, more blunt approaches, Heart Rate Variability (HRV) and Electrodermal Activity (EDA) are basic in this field [3]. These methods examine the variation in the frequencies of heartbeat and skin conductivity to measure the level of emotional activity; therefore, it is easier to associate the changes in human physiological characteristics with their emotions.

These foundation measures are followed by higher ones which are inclusive of temperature and respiration patterns and give extra information about emotions. Taking into account changes in body temperature and the rate of breathing, the proposed emotions are considered as a complex background, therefore, the efficiency of the ratings of recognizing emotions is higher.

For dimension reduction, principal component analysis (PCA) is used and, therefore, takes complexity to the next level of emotion recognition [4]. This technique helps in organizing the extensive information into more comprehensible and hence readable data. This technique prepares the ground for advanced classification models in conjunction with a technique known as artificial neural networks (ANN). These are developed using comparative experiments including Fisher + Support Vector Machine (SVM) and PCA + ANN to improve the accuracy of emotion prediction depending on different cases. The incorporation of computer vision technologies adds another perspective to the existing possibilities of emotion recognition [4]. Such methods as the pyramid of the histogram of gradients (PHOG) and the local phase quantization (LPQ) are used for describing the shapes and the appearances of human faces [5]. These features are obtained from critical frames, selected using K- Means clustering performed on normalized shape vectors formed based on Constrained Local Model (CLM)-constrained face tracking. This method focuses on analyzing the emotions that are displayed in the facial expressions most informatively, thus enhancing the emotion recognition system's efficiency in recognizing emotions based on the facial expressions.

The Emotions in Context (EMOTIC) database extends the advances of the research further by including facial expressions but training Convolutional Neural Networks (CNNs) to recognize not only the person but also the environment [6]. This approach underlines the influence of the context in the recognition of the emotions and offers a holistic reference framework making it possible to classify the emotional status in the real world rather than in controlled conditions. The purpose of this research is to review the state of existing technologies for automated emotion evaluation (AEE) and analyze their applicability in specific areas including robotics, marketing, education, and entertainment [6,7]. Consequently, this research delivers a systematic literature review of Automated Emotion Recognition (AER) and its concepts and applications, investigators' core technologies and methodologies, and assesses the performance of critical technologies in emotion measurement and its potential to affect the quality of human-machine interaction in IoT and affective computing to further intelligent empathetic machine creation.

2. Methodology

2.1. Dataset Description and Preprocessing

The FER2013 database is used in this research on Facial Emotion Recognition utilizing CNNs for this study. This dataset consists of 35,887 samples of which each is a 48 x 48-pixel grayscale image of the face expressing different emotions [8]. These expressions are categorized into seven distinct emotion categories: It consists of seven basic emotions namely anger, disgust, fear, happiness, sadness, surprise, and neutrality. Each entry in the dataset includes three key components: the emotion, pixel data for the face, and use data to specify the pixel data for training, public test, or otherwise [8].

DOI: 10.54254/2754-1169/135/2024.18696

This is beneficial for the structured format as allows for systematic training, validation, and testing within machine learning frameworks.

This makes the FER2013 dataset unique for building and evaluating advanced systems for affect recognizing with still few images. Closely related to deep learning, this research utilizes CNNs that can effectively learn and extract the most relevant patterns from the pixel layers in an image to achieve accurate classification of the described emotional states. The large span of expressions for each demographic increases the realism of the data and improves the effectiveness and feasibility of emotion recognition in practice.

2.2. CNN

For the emotional recognition of the EEG signals, this work makes use of the CNN, which is an enhanced form of the neural network architecture that has been designed for the processing of complicated signal data [8]. CNNs, especially, well-known due to their deep learning functions, are used for analyzing spatial hierarchies of data and involve convolutional layers and pooling layers.

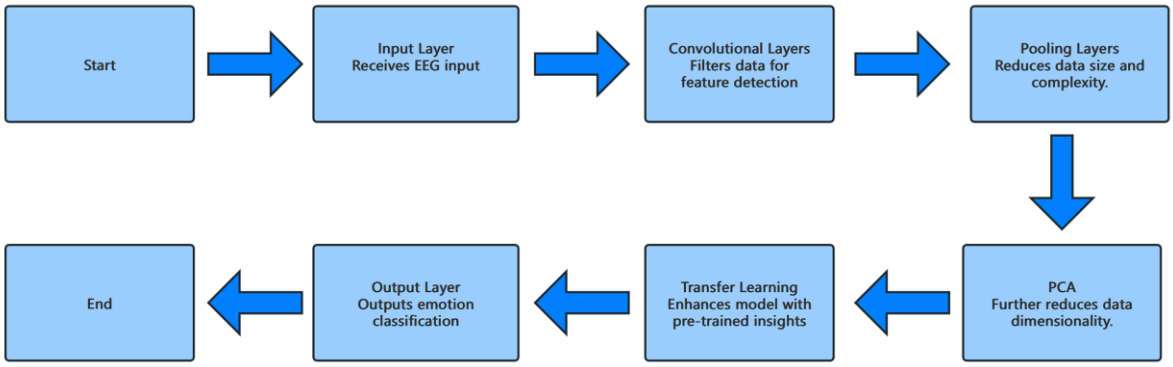

Figure 1: Architecture of CNN.

In this research, the CNN architecture (see in Figure 1) is specifically constructed based on the need for extracting and classifying the emotions based on the electroencephalogram (EEG) signals [4]. Increasingly, the first layer of the CNN utilizes several filters to conduct convolutions on the entered EEG input data to extract primary features at different scales and orientations [4]. These convolutional layers are essential for the identification of fine features that point towards different emotional conditions.

After the convolutional layers, there are the pooling layers that bring down the size of the data whilst proceeding to maintain the significant feature information. This reduction apart from enhancing the computational aspect also aids in minimizing the overfitting problem since the input data is abstracted to complexity before getting to the subsequent layers [9]. To improve the feature extraction process, the PCR preprocessing step is thus performed to compress the EEG data before it is fed into the CNN [9]. This step also ensures that the network is mainly confined to the most important features that contain the higher predictive value of the different emotional states for the intended framework thereby making the latter more accurate and faster.

The CNN in this study also includes additional layers that are fine-tuned in the transfer learning approach, whereby a pre-trained model is trained to identify the emotional features from the EEG signals. [4] This approach relies on patterns learned from large datasets on which the model was trained and which can help to increase the model's ability to generalize from EEG data to emotion recognition [4]. In this context, transfer learning accelerates the training and enhances the model's capacity with a better understanding of human emotions through EEG. All the mentioned layers including convolutional layers, pooling layers, PCA for feature reduction, and transfer learning for

DOI: 10.54254/2754-1169/135/2024.18696

adaptation contribute to the proposed efficient and consolidated framework for emotion recognition. Such an approach helps to manage the intricacies of EEG and enables the CNN to accurately detect a large number of human emotions with references [9]. The care given in designing these layers as well as the implementation of the design reveals the model's highly evolved ability to understand the complexities of the EEG signal as well as have major implications for real-world applications, especially in areas that call for complex emotion identification. The opportunities of new products of neural network technologies [9].

2.3. The Emotion Recognition

The EMOTIC is a well-rounded database that has been developed to enhance the processes of emotion recognition with the features of the individual and the surroundings [6]. Traditional approaches to emotion recognition, which primarily focus on facial expressions, often limit the understanding of emotions to six primary categories: Anger, disgust, fear, happiness, sadness, and surprise, these are widely categorized as basic emotions [6]. However, it is important to note that the aspects of context can influence emotional perception because people's environment and behavior explain more about them. With the help of context, which is implemented in EMOTIC, the analysis is considerably more detailed and, therefore, less subjective [7].

The EMOTIC dataset includes 18,316 images which have 23,788 annotated individuals in real and natural scenes [6]. It features two annotation systems: one that consists of 26 single emotional types and the second, that of differential emotional values, including Valence, Arousal, and Dominance (VAD). The discrete categories are a broad array of feelings of interest, desire, and burnout and offer concrete names of all the feelings [7]. Compared to categorized dimensions, the continuous dimensions provide a smoother generalization of emotions; Valence translates the positive or negative aspect of an emotion, Arousal determines the level of aggression or serenity of an emotion, and finally, Dominance establishes how much control one has over an emotion. The distribution of the dataset is into the training set (70%), validation set (10%), and test set (20%), and the annotations are gathered through Automated Mechanical Transmission (AMT) [7]. Employees classified photographs by the following discrete emotions and VAD dimensions, as well as additional features such as the age and gender of the person depicted in the picture [6]. Therefore, the main advantage of the dataset is the integration of these systems for the improved and more complex emotion recognition model.

The first and foremost aim of the proposed EMOTIC is to teach a CNN-based model to understand the emotion of an individual in the image along with the context of that image [7]. The model's architecture has three key components: feature extractors and an FPN as well as two fusion networks. The first feature extractor is trained to focus on the person in the image while the second is trained to view the entire image to capture other global [6,7]. The fusion network in turn utilizes such extracted features for estimating both the discrete emotional categories, as well as the continuous dimensions.

It also incorporates the feature extraction modules that are truncated low-rank filter CNNs which is computationally efficient with high accuracy as stated in [7]. These extractors work on the grounded features of the person and aerial features of the scene. A global average pooling layer decreases the number of features and the next layer is a fully connected layer that produces a 256-element vector [6]. This vector is then processed by two separate branches: The first one is for predicting the discrete emotional categories and the second one is for the continuing dimensions of the identified emotions. The training process uses a joint loss function the goal of which is to obtain the best discrete categories as well as the best continuous dimensions. The discrete categories are trained under a weighted Euclidean loss to address the problem of class skew where certain categories are less frequent than others [6]. The involved continuous dimensions are optimized with a Euclidean loss with an error margin incorporated because of the possibility of subjectivity in the labeling done by people.

DOI: 10.54254/2754-1169/135/2024.18696

Incorporating the ImageNet and Places pre-trained models causes an enhancement in the performance of the trained model [7]. As for the metric of ImageNet, it aids in determining the content of the person's image region while Places improves the context of the scene. Such datasets enable the model to get information about individual and contextual tendencies that are connected with emotions [6]. After the training of the model, the performance of the model is tested on the testing data, and from the results, one can find that the fusion of the person and the scene yields the best results for recognizing the affective states [7]. For many of the emotional categories, including engagement, excitement as well as happiness, the inputs, in this case, the body and the image (context), yield a higher precision. The continuous dimensions (Valence, Arousal, Dominance) are also well predicted, with low error rates, to suggest that the model is an effective means of understanding the finer details of the emotions that are being experienced [7].

For instance, the EMOTIC model does not require a subject's face to be present to identify each person's emotion; it also defines the common contextual cues and body language [6]. This is especially important for real-world applications where faces may be partially occluded or non-frontal. The system even works when the face is partially visible or when a person turns away and that makes it suitable for just any occasion.

2.4. PHOG and LPQ

The method introduced for automatic emotion recognition and presented for the FERA 2011 competition jointly uses shape and appearance features for the classification of emotion from facial images in image sequences. The system takes the Pyramid of PHOG and LPQ features to identify the face structure and texture [5]. The process starts with face tracking using CLM that give details of facial points within image sequences. This tracking results in shape vectors that are normalized to which K-means clustering is combined to select keyframes [5]. These are commonly referred to as keyframes and they are employed in feature extraction having been extrapolated from crucial instances of facial expression transformation.

In this way, PHOG features are derived from the selected keyframes where the face is divided into spatial subregions, and the edges' orientation at various pyramid levels is assessed. This method enables a capture of the structural form of the face required in the recognition of different facial expressions [5]. In the LPQ case, LPQ features are exploited to extract appearance information by reconstructing the local phase of the facial texture. In detail, LPQ is more sensitive in outlining facial features including wrinkles or even slight changes in texture play an important role in the identification of emotions [5]. Altogether, PHOG teamed up with LPQ can give a full description of the face features in terms of shape and texture [10].

Once the PHOG and LPQ features are extracted, the system employs two classification methods: SVM and a recently developed algorithm, the Largest Margin Nearest Neighbor (LMNN) [5]. As it regards the division of the different categories of emotions, it employs SVM to search for the best decision surface. Also, for the process of emotion classification, LMNN learns the distance metric of separation of other classes' data points is optimized. The performance of the system was evaluated on this Association of Enterprises for the Study of Radon Emission and its Pollutant Effects (GEMEP-FERA) database in which several sequences they recorded actors feeling emotions.[10]. It results from the creation of partitions for person-specific and person-independent to perform an adequate evaluation of the system. The method was tested in three prototypes. Initially, only PHOG features combined with SVM for classification have been developed with the first configuration [10]. This approach gave moderate accuracy and this was because it only depended on the shape information of the face. The second configuration coupled both PHOG and LPQ features with SVM, which brought substantial changes in performance [5]. As can be seen from the experiments, incorporating LPQ features enabled the system to obtain richer appearance characteristics to facilitate emotion

DOI: 10.54254/2754-1169/135/2024.18696

classification. The third configuration used PHOG and LPQ features with LMNN for classification [5]. This method was deemed to be slightly better than the earlier introduced SVM-based approach, especially in the person-specific tests but suffers more in the person-independent data.

As for outcomes, about person-specific as well as person-independent tasks, the harmonic mean of accuracy was the highest for the integration of PHOG and LPQ features with SVM [5]. In certain cases LMNN performance was better, however, it was not significantly better than SVM [5]. In general, the system examined the necessity of integration of shape and appearance features for emotion modeling and specified that applying of K-means algorithm for the selection of the keyframes minimizes the amount of the input data in the image sequences.

3. Result and Discussion

3.1. Results Analysis

More specifically, this work focuses on the CNNs for emotion recognition through analysis of the EEG signals [4]. The EEG data collected were first preprocessed and then reduced in dimensionality using PCA so that the CNNs could learn to extract important emotional parameters on their own. When compared with other classification models such as KN-5NN & SVM, it was observed that CNN performed much better with an accuracy of 84%. 5 % for valence classification and 81 The identification accuracy is equally determined to be 2% for arousal classification which is remarkably better than previous techniques [4]. The process of PCA helped here a lot, as it cut the number of features from 8064 to 40, which improved the model's performance and effectiveness not losing crucial information in the process.

Given that the EEG data was reshaped into a 2D array similar to image data, convolutional and pooling layers CNN models were able to classify the data successfully [8]. The model also showed robustness against noise, which presented very good results even when the classifier had to handle noisy or relatively more challenging to classify EEG data due to external factors or inter-subject variations. In addition, there were even denser models including Densely Connected Convolutional Network (DenseNet)-161, which also demonstrated excellent performance in grasping multi-emotional cues and non-frontal vision with an accuracy level of 96%. 93% on the DriverNet dataset; 51% on the karolinska directed emotional faces (KDEF) dataset and 99. 52% relatively on the Japanese Female Facial Expression (JAFFE) dataset. The second advantage was that of transfer learning (TL), this is because as suggested in, Visual Geometry Group (VGG) 16 and DenseNet-161 pre-trained models were retrained for emotion recognition [8]. This had a great impact on reducing the need for big data and the resources required to handle them. In the pipeline strategy which was applied when cascading these models, overfitting was reduced meaning the useful information acquired in the previous layers would be retained. CNNs thus proved that they have a good generalization capacity as they performed well on different datasets and in frontal and profile poses hence making them suitable for use in emotion recognition in real-world applications [4].

Thirdly, CNNs are also highly extensible, they can process large EEG data sets by concentrating on local input areas to perform convolutions. This is especially the case when working with the spatial and temporal characteristics of various EEG signals. CNNs also well fit other preprocessing techniques such as reducing the dimensions and data augmentation, and thus CNNs can pay more attention to salient features regardless of the noise or missing data from the images. The inherent ability to handle change processes normally existing in images including rotation, flipping resizing, etc., was transferred to EEG data to ensure good performance regardless of the level of change [8]. Hence, CNNs outperformed in processing continuous and high-resolution signals in real-time applications such as Better Cotton Initiative (BCI) or a healthcare system, which makes CNN suitable for real-time emotion recognition [4].

DOI: 10.54254/2754-1169/135/2024.18696

3.2. Discussion

This study shows that CNNs are effective, especially in esteeming emotions when used with PCA for data reduction and transfer learning. CNNs' automatic feature extraction function lessens the amount of feature engineering needed, greatly improving outcomes in emotion regulation [4]. Single-shot methods such as DenseNet-161 were found to be particularly useful in identifying the multi-layered emotional states, which was several percentage points higher than the baseline. Transfer learning was then used to enhance performance from a reduction in large datasets and excellent generalization traits of CNNs in different datasets. For example, in the real experiment, the accuracy of the models was higher than 96%, on the KDEF and JAFFE [8].

However, CNNs do have limitations. While PCA was effective for improving efficiency, there is a risk of losing critical emotional information during dimensionality reduction, which may hinder the model's ability to classify complex emotional states. Furthermore, the deep architectures used in CNNs, while powerful, require substantial computational resources, potentially limiting their use in resource-constrained environments [4]. Despite the advantages of transfer learning, the model's generalization capability, particularly for specific emotional categories, could be further improved. Additionally, PCA's assumption of linear separability might not hold in more complex, non-linear emotion recognition tasks, limiting its effectiveness [4]. Future research could explore more advanced feature selection techniques or hybrid approaches to improve CNN performance in emotion recognition. In conclusion, CNNs have demonstrated remarkable performance in emotion recognition tasks, particularly when processing EEG signals [10]. However, addressing the limitations of dimensionality reduction and computational demands, as well as further improving the model's generalization capabilities, will be critical for broader applications in real-world emotion recognition systems.

4. Conclusion

This study offers a thorough evaluation of three key emotion recognition techniques: CNNs, the EMOTIC dataset, and the feature extraction techniques described as follows: PCA and LPQ. All of these methods have their advantages – CNNs apply features from EEG and image information with no preceding preprocessing, PCA helps to improve model efficiency by reducing its dimensions, while LPQ is considered to analyze texture effectively in conditions of guaranteed lighting. The use of both techniques ensured high accuracy in grouping emotions especially when CNNs were used on EEG data. The EMOTIC dataset was also shown to help identify different degrees of emotions as it is limited to only seven primary ones and covers both the face and the body. The LPQ again enhanced the result of the facial recognition tasks in the unconstrained environment. However, there is still a lot to do, especially in the areas of real-time data analysis and using the results for different types of data. To strengthen the proposed method for future work, more modalities such as audio and physiological signals should be included while other work should be directed towards surpassing the simplicity of data by developing temporal models for capturing emotion trends within a particular period. These enhancements may result in higher accuracy and expansion of emotion detection for practical applications in health, society, and human-computer interaction.

References

[1] *Finkelstein, V. (1996) Outside, inside out. Coalition, 30-36.*
[2] *Juckel, G., Heinisch, C., Welpinghus, A., & Brüne, M. (2018) Understanding another person's emotions—an interdisciplinary research approach. Frontiers in psychiatry, 9, 414.*
[3] *Egger, M., Ley, M., & Hanke, S. (2019) Emotion recognition from physiological signal analysis: A review. Electronic Notes in Theoretical Computer Science, 343, 35-55.*

DOI: 10.54254/2754-1169/135/2024.18696

[4] *Chen, L., Mao, X., Xue, Y., & Cheng, L.L. (2012) Speech emotion recognition: Features and classification models. Digital signal processing, 22(6), 1154-1160.*
[5] *Dhall, A., Asthana, A., Goecke, R., & Gedeon, T. (2011) Emotion recognition using PHOG and LPQ features. In IEEE International Conference on Automatic Face & Gesture Recognition, 878-883.*
[6] *Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017) Emotion recognition in context. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1667-1675.*
[7] *Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2019) Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence, 42(11), 2755-2766.*
[8] *Eravcı, O. (2021). Emotion Detection using CNN. Retrieved from https://www.kaggle.com/code/oykuer/emotion-detection-using-cnn?scriptVersionId=104520756*
[9] *Cao, G., Ma, Y., Meng, X., Gao, Y., & Meng, M. (2019) Emotion recognition based on CNN. In Chinese Control Conference, 8627-8630.*
[10] *Sarangi, P.P., Mishra, B.S.P., & Dehuri, S. (2019) Fusion of PHOG and LDP local descriptors for kernel-based ear biometric recognition. Multimedia Tools and Applications, 78, 9595-9623.*