2023 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom) | 979-8-3503-9341-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/MoSICom59118.2023.10458797

2D CNN vs 3D CNN: An Empirical Study on Deep Learning-based Facial Emotion Recognition

Arav Dhoot
EECE Department Intern
American University in Dubai
aravdhoot@gmail.com

N. Ben Hadj-Alouane
EECE Department
American University in Dubai
nalouane@aud.edu

M. Turki-Hadj Alouane
College of Computer Science
King Khalid University, Abha, Saudi Arabia
malouane@kku.edu.sa

***Abstract*—Human emotion detection is a significant challenge in computer-based, automatic emotion detection. Deep learning techniques have been employed to extract facial features from videos, utilizing both visual and auditory input. However, the effectiveness of deep learning methods, specifically using a lightweight model, to exclusively extract facial features from videos has not been extensively explored. Consequently, this study aimed to compare the performance of two distinct 3-dimensional convolutional neural networks (CNNs) and 2-dimensional CNNs in extracting emotion from visual input in videos. Two benchmark databases, namely the Acted Facial Expressions in the Wild database and the eNTERFACE'05 Audio-Visual Emotion Dataset, were employed for evaluation. Multiple experiments conducted during this study concluded that the 2D CNN exhibited superior performance in accurately detecting facial emotion while also utilizing fewer computational resources in comparison to the 3D CNN.**

***Index Terms*—2D convolutional neural network, 3D convolutional neural network, emotion recognition, eNTERFACE'05, Acted Facial Expressions in the Wild**

I. Introduction

Recent advancements in machine learning and deep learning techniques have brought significant progress to the field of computer-based emotion detection, revolutionizing the analysis and interpretation of emotional cues. Particularly, deep learning algorithms, which leverage intricate neural network architectures, have demonstrated exceptional capabilities in capturing complex patterns and representations from high-dimensional data. These algorithms excel in learning hierarchical features and can extract subtle emotional cues that may go unnoticed by human observers [1].

Human emotion detection has been one of the key problems that have interested researchers in the fields of machine learning and deep learning for a very long time. Accurate detection of human emotions will be pivotal in the development of the fields of smart machines, human-robot interaction, remote- and tele-health monitoring, automated call centers, gaming, and smart communication. Already, multiple media and techniques have been utilized to detect emotions. For example, emotions have been detected through text analysis [2] [3] [4], speech processing [5] [6], image analysis [7] [8], and even physiological signals like ECG [9].

Videos are one of the most common mediums to detect emotions [10]. Videos are not only more accessible than media such as physiological signals but, with advancements in smartphone technology and as a result of the Internet becoming commonplace, they have also become ubiquitous. Therefore, our objective in this paper, is to use convolutional neural networks, one of the most popular deep learning architectures, to perform video-based emotion analysis. Using different-dimensional convolutional neural networks, we hope to evaluate architectures based on their computational expense and accuracy, paving the way for future video-based emotion detection systems.

In this paper, we focus on the utilization of 2-dimensional convolutional neural networks (2D CNNs) and 3-dimensional convolutional neural networks (3D CNNs) to detect emotions from videos. Unlike previous studies that have performed video-based emotion analysis, we do not incorporate audio input from videos for classification. Instead, we solely focus on the visual cues. This enables a direct comparison between 2D CNNs and 3D CNNs in being able to extract features from images and detect emotions.

We propose and prove, with empirical evidence from extensive experimental testing on two benchmark databases commonly employed in the field of emotion detection research - eNTERFACE'05 [11] and Acted Facial Expressions in the Wild (AFEW) [12] - that 2D CNNs are more effective at extracting facial cues. Additionally, while all models utilized in this study are lightweight, in terms of their minimal computational resource requirements during the training phases, 2D CNNs exhibit superior performance while utilizing fewer computational resources than their 3D counterparts.

The remainder of this paper is organized as follows. In the subsequent section, we provide a comprehensive review of related works in the same domain. Section 3 presents and discusses our model design, highlighting the datasets employed in our experiments. In Section 4, we delve into a detailed analysis of the results obtained. Finally, Section 5 concludes our study and outlines potential avenues for future research.

II. Related Works

The use of deep learning architectures, specifically, neural networks, is very common for the detection of emotion [13]. Advanced neural networks have been used to extract features from audio data to detect emotion [14] [15]. The use of convolutional neural networks, which is the primary deep

learning architecture being discussed in this paper, is more common in extracting visual cues from still images and videos.

In most studies where an image is used as the only input, a 2D CNN is preferred [16]. 2D CNNs display great versatility. In fact, a plethora of different CNN models has been developed for different emotion recognition tasks, including the fusion of several sub-models [17]. Face detection and image cropping are also very common in image-based emotion recognition. This helps remove noise from the images to point the focus on the face [18]. For image-based emotion recognition, a few popular image-based databases that were often consulted include Facial Expression Recognition 2013 database [19], the Extended Cohn-Kanade database [20], and the Japanese Female Facial Expression database [21].

In studies where a mixture of audio and visual input is used for emotion recognition, a 3D CNN architecture is preferred for the visual input. However, 2D CNNs are preferred, because of their lower computation usage [22]. When dealing with audio-visual input, the separation of audio and video modalities is important. Subsequently, the fusion of results from both modalities is also essential for the ultimate classification. There are four primary methods to fuse results from both modalities: feature-level fusion, decision-level fusion, score-level fusion, and model-level fusion [23].

Unlike image-based emotion recognition, frame selection is an important aspect of video-based emotion recognition. There are a few different techniques used for frame selection: splitting video clips into smaller uniform segments [23], splitting the entire video into frames [24], or computationally selecting a few keyframes per video [25] [26], which is the technique used in our study. Additionally, works referred to above cropped video frames to only include faces to reduce unnecessary information that might interfere with the neural networks' computations. The face detection algorithm of choice was the Viola-Jones face identification algorithm, also known as the Haar cascade algorithm [27].

It is also common, when using video data, to train neural networks on multiple databases, to increase the robustness of the model [25].

Overall, the most commonly used databases for video-based emotion recognition include the eNTERFACE'05 Audio-Visual Dataset and the Acted Facial Expressions in the Wild database, which were both the focus of our study. Additionally, the Surrey Audio-Visual Expressed Emotion (SAVEE) database [28] was also considered by multiple studies. We didn't consult the SAVEE dataset in our study because it is limited to only four British male subjects. As a result, training our models on the SAVEE dataset would have introduced a gender, and possibly, geographical bias, making it ineffective for multi-source training. Furthermore, video recordings in the SAVEE database are accompanied by facial key points. However, this scenario is difficult to recreate with other real-time videos.

In addition, it is important to note that [25] is a comprehensive study in the domain of audiovisual emotion detection using deep learning techniques. We reference the same databases and we share similar video-processing techniques in our study. Hence, the study is widely referenced throughout our paper. Further, note that most of the neural network architectures used by studies referenced above are computationally expensive, including 2D CNNs. For example, [25] uses a deep residual network that is 50 layers deep.

III. Data-sets and Model Design

We start with a succinct description of the important features of the challenging databases we chose as the basis of our paper. We then move to describe the models we propose to use for emotion detection.

A. Datasets Description

For our paper, we consulted two popular video-based databases that are widely used in machine learning and deep learning for detecting emotions: eNTERFACE'05 [11] and Acted Facial Expressions in the Wild (AFEW) [12].

1) eNTERFACE'05 Audio-Visual Emotion Dataset: The eNTERFACE'05 Audio-Visual Emotion Dataset comprises 1166 video clips, of which 77% were clips featuring men, while the remaining 23% were clips featuring women. The recordings were presented by 42 subjects from 14 different nationalities and represented the six important emotions: anger, disgust, fear, happiness, sadness, and surprise. For each emotion, the subjects were expected to produce 5 recorded reactions to a given situation. All videos underwent review by two experts, and any videos that failed to elicit the expected emotion were discarded. Additionally, all videos in this database were recorded in a controlled recording room. Hence, all videos captured frontal views as shown in Figure 1.

Fig. 1. Frames from a random selection of videos from the eNTERFACE'05 database.

2) Acted Facial Expressions in the Wild Database: The second database used for this study was the Acted Facial Expressions in the Wild (AFEW) database, which comprises video clips from movies. The database encompasses 957 videos distributed across seven emotional states, in addition to the six common states mentioned above, we also find neutrality used as an emotion. The videos in this database have been gathered from diverse scenarios, including conditions with low lighting and videos featuring multiple people in the same frame. Consequently, the AFEW database is a better reflection of real-world conditions. In fact, it is always presented in the literature as the most challenging database for emotion detection [29]. Figure 2 shows some frames samples from a random selection of videos from the AFEW database

Fig. 2. Frames from a random selection of AFEW database videos.

There are notable distinctions between the two databases. Firstly, the eNTERFACE'05 database consists of six emotion classes, while the AFEW database encompasses 6 emotion classes and an additional neutral class. Hence, to ensure class commonality between the AFEW and the eNTERFACE'05 databases for multi-source training, for the purpose of this study, the neutral class, with its corresponding videos, was disregarded from AFEW. Secondly, the video subjects in the AFEW database, span a larger age spectrum, ranging from 1 year to 70 years. Thirdly, the average length of video recordings in the eNTERFACE'05 database is greater than the average length of video recordings in the AFEW database. Finally, the videos in the AFEW database represent a wider range of conditions and are closer to real-life emotion detection conditions. In contrast, the videos in the eNTERFACE'05 database have been shot in very uniform, lab-like conditions.

By using both databases we offer to our study both ends of the spectrum in emotion detection: the standard lab-like data and the realistic data.

B. Model Design

For the purpose of this paper, a lightweight 2D CNN with three convolutional layers is employed. The CNN architecture consists of two consecutive convolution layers, each playing a crucial role in feature extraction. The first convolution layer focuses on extracting the low-level features, such as edges and textures, from the input frames. This initial feature extraction allows the model to capture fundamental patterns and forms the foundation for subsequent analysis. The second convolution layer builds upon the low-level features obtained from the previous layer, extracting higher-level features that represent more complex patterns. By incorporating this additional layer, the model gains the ability to capture more advanced and abstract patterns present in the data. The combination of these consecutive convolution layers facilitates a hierarchical feature learning process, enabling the model to progressively extract and analyze increasingly sophisticated representations.

To ensure a fair evaluation and comparison between the 2D and 3D architectures, the same configuration is used for both models, as shown in Figure 3. This uniformity guarantees a consistent basis for judging the performance and capabilities of each architecture. By employing this standardized setup, any observed differences in performance can be attributed solely to the architectural variances between the 2D and 3D CNNs. It is important to note that while the layers in both the models are the same in terms of the function they perform, they differ in dimensionality.

It must also be noted that in our study, since our objective lies in primarily understanding the differences in the performance of 2D and 3D CNNs, we use shallow models. We expect further work in the field that builds on this research to build deeper models that are better at extracting the visual features from video frames.

C. Experimental Setup

For the purpose of our study, a meticulous process was undertaken to select a representative set of frames from each video for analysis. Initially, we solely followed the frame selection procedure outlined in [25]. However, during the implementation of this method, two significant challenges were encountered. Firstly, frames were not consistently detected for certain videos. For some videos, no frames were detected at all. Hence, using the procedure mentioned in [25], some videos in the database would not have been represented at all, leading to incomplete data that could potentially introduce bias or hinder accurate analysis. Secondly, the number of frames detected per video exhibited variability, rendering it incompatible with the requirements of a 3D CNN that requires the same number of frames input for each video.

To overcome these challenges and ensure a consistent input for our 3D CNN model, a modified frame selection approach was devised. First, we used the method adopted by [25]. Then, empirically, we noticed most videos usually contained more facial cues that might help with emotion detection toward the end of the video. Thus, in addition to the frames selected using the methods of [25], we added the last frames from each video,

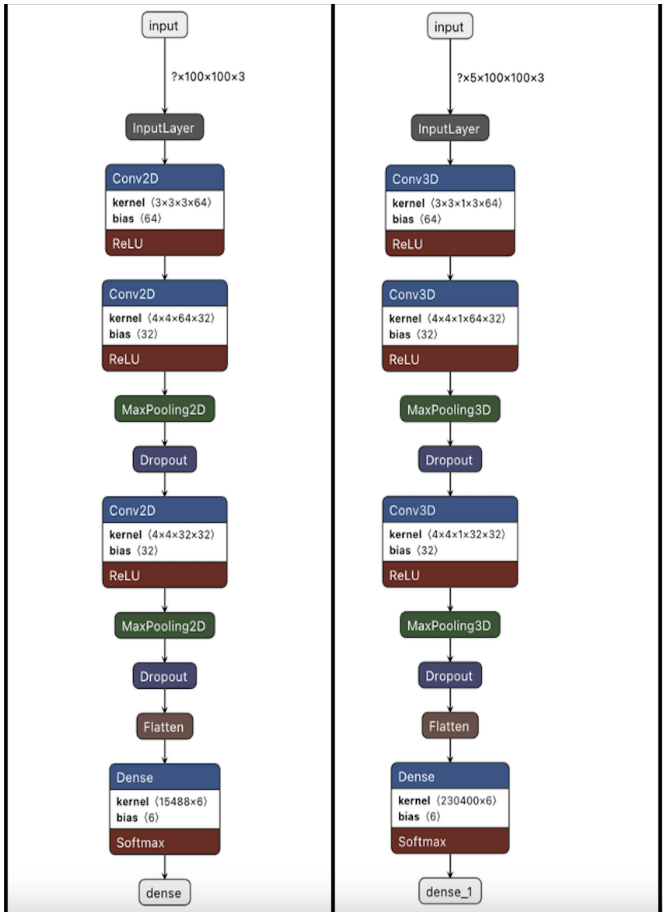

Fig. 3. 2D CNN architecture design (left). 3D CNN architecture design (right)

until a total of five frames per video were accumulated. Hence, a total of five frames per video were fed into the 3D CNN at once. This methodology not only resolved the issue of missing frames but also provided a uniform input per video. We had 5 frames of input per video that served as a single input for the 3D CNN. For the 2D CNN, on the other hand, each frame served as a single input.

The number of images in the training, validation, and testing sub-databases for both databases has been reported in Table I.

TABLE I
SIZES OF THE TRAINING, VALIDATION, AND TESTING SUB-DATABASES

	AFEW	*eNTERFACE'05*
Training data size	4416	2260
Validation data size	620	445
Testing data size	1258	220

After the frame selection process, a bounding box was applied to localize and extract facial regions from each video. To accomplish this, we employed the Viola-Jones face identification algorithm [27]. The Haar cascade identifier first divides the image into a grid of small windows, then calculates a set of Haar features for each window. These features are simple mathematical functions that measure the difference between the brightness of different parts of the window. In the AFEW database, it was observed that certain videos contained multiple faces. In such scenarios, a decision was made to select the largest face among the detected faces based on the area of the bounding box.

Table II shows the values of the hyperparameters used for training.

TABLE II
HYPERPARAMETERS USED FOR TRAINING

	AFEW	*eNTERFACE'05*
Batch Size	32	32
Learning Rate	1e-5	5e-4

The same values of the hyperparameters were maintained across the 2D CNN and the 3D CNNs. The NVIDIA Tesla T4 GPU was used for training the 2D CNN, and the NVIDIA Tesla V100 GPU was used for training the 3D CNN.

D. Results

To evaluate the performance of each database individually, the database was divided into train, test, and validation sub-databases in a 7:2:1 ratio. Three different experiments were conducted, one with a 2D CNN with color, one with a 3D CNN grayscale, and one with 3D CNN with color. In the second experiment, the frames were converted to grayscale in an effort to reduce computational resources.

Overall, the 2D CNN consisted of 143,942 parameters, while the 3D CNN for grayscale consisted of 1,433,414 parameters. On the other hand, the 3D CNN for color consisted of 2,317,254 parameters. Hence, the 3D CNNs use approximately 10 times and 16 times more computational resources than the 2D CNN respectively. Additionally, the 3D CNN, which is trained on colored images, uses approximately 1.6 times more computational resources than the 3D CNN, which is trained on grayscale images.

In all cases, the models were trained on 200 epochs. The obtained results were reported in Table III

TABLE III
THE TEST ACCURACIES OBTAINED FOR BOTH DATASETS

Database	*2D CNN*	*3D CNN (no color)*	*3D CNN (color)*
eNTERFACE'05	67.7%	59.6%	61.4%
AFEW	29.7%	19.7%	20.1%

As reported in Table III, it is obvious that the 2D CNN performs better than both 3D CNNs. However, the accuracy of all three models is close to the level of random guessing when tested on the AFEW database. Further, it is noted that the 3D CNN performs slightly better when the images retain their color. However, the improvements are negligible considering the significant increase in the amount of computational resources required to train the model.

In Table IV the results from the experiments conducted in [25] have been reported. In [25], a 50-layer 2D CNN was used for training. In addition, they had 1570 images and 522 images for the eNTERFACE'05 and AFEW databases respectively.

TABLE IV
RESULTS REPORTED IN [25]

eNTERFACE'05	*AFEW*
50.2%	46.6%

In the context of these scores, it is important to note that we had a significantly larger dataset that contained images from a wider variety of situations. We achieved a significantly lower score on the AFEW dataset compared to the results reported by [25]. More interestingly, we achieved a significantly higher score on the eNTERFACE'05 database on all three models, even while having a shallower model. Having a larger data size on the eNTERFACE'05 significantly helped the models to learn a wider range of features. In the same vein, having a larger data size on the AFEW database, which represents a wider range of features, led to the lower performance of models. Having a wide range of features and a small model, made it difficult for the model to effectively learn the features.

To gain further insights into the biases encountered during the training process of our model, we conducted another so-called blended training experiment that used inputs from both databases for training at the same time. In this experiment, we merged the training and validation sub-databases from AFEW and eNTERFACE'05. During the merged training, we employed a learning rate of 5e-4 and utilized a batch size of 32. For this experiment, we only trained the 3D CNN exclusively on images that contained color information. The models were trained for 200 epochs using the merged training database. Subsequently, the trained models were evaluated on separate testing sub-databases. The results obtained from this blended training experiment are presented in Table V.

TABLE V
TEST ACCURACIES OBTAINED WHEN MODELS WERE TRAINED ON MERGED TRAINING DATASETS.

Database	*2D CNN*	*3D CNN*
eNTERFACE'05	16.1%	16.5%
AFEW	17.3%	17.2%

Blended training produced sub-par testing results on both databases. In fact, on the eNTERFACE'05 database, both models produced results worse than random guessing. Regardless, blended training reveals interesting results. Firstly, it becomes apparent that the features represented by the AFEW database and the eNTERFACES'05 database are conflicting. This ends up resulting in poor performance on the test database. Secondly, as a result of the diverse range of situations the videos in the AFEW database were taken in, the videos in the database represent stronger features compared to the eNTERFACES'05 database. As a result, even with a smaller training database size than the eNTERFACES'05 database, AFEW achieves a higher performance score. More notably, the scores achieved by the AFEW database on blended training are closer to the score it received when it was trained solely on AFEW.

IV. DISCUSSIONS

In this section, we delve into two primary aspects of our study. Firstly, we address the lower accuracy scores we obtained on the AFEW database compared to [25]. Secondly, we explore the reasons behind the superior performance of the 2D CNN over the 3D CNN in our experiments.

Two factors contribute to a lower accuracy score on the AFEW database in our study compared to the results reported in [25]. Firstly, the disparity arises from the difference in the frame selection methodology employed. Through the methodology adopted by [25], a frame is selected per video, at maximum, while other videos in the database are disregarded entirely. This approach results in an incomplete representation of the database. In contrast, our study utilizes a combination of the aforementioned method along with the selection of the final frames from each video until a total of five frames per video is obtained. Consequently, our approach results in the complete representation of the database, albeit with the tradeoff of including frames that might not be the most optimum.

Furthermore, it is noteworthy that the model used in [25] utilizes a significantly deeper network architecture than the one employed in our study. Despite the complexity and increased depth of their model, requiring far more computational resources, our research achieves a higher accuracy score on the eNTERFACES'05 database.

Regarding the second point of the discussion, we delve into the reasons behind the superior performance of the 2D CNN over the 3D CNN in our experiments. Firstly, the 3D CNN takes multiple frames of input simultaneously to incorporate temporal data. In our case, we feed five frames from a video. These five frames constitute a single input. In contrast, the 2D CNN treats each frame individually, resulting in five separate inputs. Consequently, the 2D CNN enjoys a larger training input size, despite sharing the same absolute data size.

Secondly, the larger training input size afforded to the 2D CNN allowed the model to develop a more profound understanding of feature representations compared to the 3D CNN. The increased data diversity and quantity for the 2D CNN likely contributed to its ability to generalize better on the emotion recognition task, reinforcing the performance of the 2D CNN architecture. It might be helpful to increase the size of the data to improve the performance of the 3D CNN. This would also result in an increase in the data that is available to 2D CNNs. Therefore, the 2D CNN might still outperform the 3D CNN.

Finally, the frame selection technique implemented in our study did not guarantee that the selected frames were consecutive, leading to a lack of temporal linearity in the input data for the 3D CNN. This may have hindered the model's ability to effectively capture temporal patterns and correlations within the video sequences. On the other hand, the 2D CNN

could independently process each frame, potentially mitigating the impact of non-consecutive frames and enabling it to learn meaningful features.

V. Conclusion

In this paper, we conducted an empirical study comparing the performance of 2D CNNs and 3D CNNs in the task of facial emotion recognition. We focused on video-based emotion analysis using visual cues and excluded audio input to directly compare the effectiveness of the architectures of 2D CNNs and 3D CNNs in extracting facial features and detecting emotions. Through experimentation on two benchmark databases, the eNTERFACE'05 Audio-Visual Emotion database and Acted Facial Expressions in the Wild database, we obtained valuable insights into the capabilities of 2D CNNs and 3D CNNs. Our results demonstrated that 2D CNNs outperformed 3D CNNs in terms of accuracy while utilizing fewer computational resources compared to 3D CNNs.

Our experiment results suggest that for facial emotion recognition tasks based on visual cues, 2D CNNs are more effective and efficient compared to 3D CNNs. The hierarchical feature learning capability of 2D CNNs allows them to extract facial features and capture complex patterns, enabling accurate emotion detection. Moreover, the reduced computational resources required by 2D CNNs make them a more practical choice for real-time applications or resource-constrained environments for real-time emotion recognition.

Future Work: We think further studies can delve deeper to see whether our results generalize to video evaluation using deep learning in other domains as well. Additionally, in the future, models can be improved to be able to detect a wider range of emotions. However, that might require the creation of a more detailed database. On top of that, it might be helpful to explore whether training the 3D CNN on a larger data size positively impacts its performance in video-based emotion detection. Finally, research conducted in the future can explore whether 2D CNNs continue to exhibit better performances compared to 3D CNNs even when they are made deeper.

References

[1] X. Liu, S. Li, and M. Wang, "Hierarchical attention-based multimodal Fusion Network for Video Emotion Recognition," Computational Intelligence and Neuroscience, vol. 2021, pp. 1–11, 2021.

[2] Matthew Purver and Stuart Battersby, "Experimenting with Distant Supervision for Emotion Classification," In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 482–491, 2012.

[3] R. Xu, T. Chen, Y. Xia, Q. Liu, B. Liu, and X. Wang, "Word embedding composition for data imbalances in sentiment and emotion classification," Cognitive Computation, vol. 7, no. 2, pp. 226–240, 2015.

[4] N. Rey-Villamizar et al., "Analysis of anxious word usage on online health forums," Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016.

[5] F. Yu, E. Chang, Y.-Q. Xu, and H.-Y. Shum, "Emotion detection from speech to enrich multimedia content," Advances in Multimedia Information Processing — PCM 2001, pp. 550–557, 2001.

[6] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," Proceedings of the 22nd ACM international conference on Multimedia, 2014.

[7] A. Jaiswal, A. Krishnama Raju, and S. Deb, "Facial emotion detection using Deep Learning," 2020 International Conference for Emerging Technology (INCET), 2020.

[8] I. Lasri, A. R. Solh, and M. E. Belkacemi, "Facial emotion recognition of students using convolutional neural network," 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), 2019. doi:10.1109/icds47004.2019.8942386

[9] F. Agrafioti, D. Hatzinakos and A. K. Anderson, "ECG Pattern Analysis for Emotion Detection," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 102-115, 2012.

[10] J. M. Garcia-Garcia, V. M. Penichet, and M. D. Lozano, "Emotion detection," Proceedings of the XVIII International Conference on Human Computer Interaction, 2017.

[11] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The Enterface'05 Audio-Visual Emotion Database," 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006.

[12] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Acted Facial Expressions in the Wild Database. Technical report, Australian National University, 2011.

[13] S. M. S. Abdullah, S. Y. Ameen, M. A. M. Sadeeq, and S. Zeebaree, "Multimodal Emotion Recognition Using Deep Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 02, pp. 52–58, 2021.

[14] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using Deep Neural Networks," 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), 2015.

[15] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," 2009 IEEE Workshop on Automatic Speech Recognition Understanding, 2009.

[16] C. Pramerdorfer and M. Kampel, "Facial expression recognition system using convolutional neural networks," International Journal of Recent Technology and Engineering, vol. 8, no. 2S4, pp. 603–607, 2019.

[17] A. Jaiswal, A. Krishnama Raju, and S. Deb, "Facial emotion detection using Deep Learning," 2020 International Conference for Emerging Technology (INCET), 2020.

[18] H. Boubenna and D. Lee, "Image-based emotion recognition using evolutionary algorithms," Biologically Inspired Cognitive Architectures, vol. 24, pp. 70–76, 2018.

[19] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

[20] P. Lucey et al., "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010.

[21] M. Lyons, M. Kamachiand J. Gyoba, "Japanese Female Facial Expression (JAFFE) Database," 2017.

[22] S. Jaiswal and G. C. Nandi, "Robust real-time emotion detection system using CNN architecture," Neural Computing and Applications, vol. 32, no. 15, pp. 11253–11262, 2019.

[23] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 3030–3043, 2018.

[24] Z. Farhoudi and S. Setayeshi, "Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition," Speech Communication, vol. 127, pp. 92–103, 2021.

[25] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," Machine Vision and Applications, vol. 30, no. 5, pp. 975–985, 2018.

[26] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," Information Fusion, vol. 49, pp. 69–78, 2019. doi:10.1016/j.inffus.2018.09.008

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of Simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[28] W. Wang, Machine Audition: Principles, Algorithms, and Systems. Hershey, PA: Information Science Reference, 2011.

[29] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-VA database for Valence and arousal estimation in-the-wild," Image and Vision Computing, vol. 65, pp. 23–36, 2017.