Emotion Recognition from Audio-Visual Information based on Convolutional Neural Network

Syrine Haddad
Laboratory of Robotics, Informatics and Complex Systems (RISC Lab-LR16ES07)
National Engineering School of Tunis, University of Tunis El Manar, 1002 Tunis, Tunisia
syrine.haddad@enit.utm.tn

Olfa Daassi
Laboratory of Robotics, Informatics and Complex Systems (RISC Lab-LR16ES07)
National Engineering School of Carthage, University of Carthage 2035 Carthage, Tunisia
olfa.daassi@yahoo.com

Safya Belghith
Laboratory of Robotics, Informatics and Complex Systems (RISC Lab-LR16ES07)
National Engineering School of Tunis, University of Tunis El Manar, 1002 Tunis, Tunisia
safya.belghith@enit.utm.tn

2023 International Conference on Control, Automation and Diagnosis (ICCAD) | 979-8-3503-4707-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCAD57653.2023.10152451

***Abstract*— In recent years, emotion recognition becomes a heavily researched field to consider in any project related to affective computing. Due to the almost limitless applications of this new discipline, the development of emotion recognition systems has emerged as a lucrative opportunity in the corporate sector. Emotion recognition can be detected in many ways, face, speech, text, gestures, etc. This paper proposes an emotion recognition approach using multiple information sources. Our approach involves both visual and speech information and it is based on Convolutional neural networks. The experiments prove the effectiveness of the proposed approach and the importance of combining visual with audio data.**

Keywords*— Emotion detection; audio-visual emotion detection; Convolutional Neural Network; multimodal emotion recognition.

I. INTRODUCTION

Emotions are present in almost every decision and every moment of our lives. Recognizing emotions is intriguing because we could interact more effectively if we knew the other person's feelings.

To improve the user experience, researchers are trying to facilitate the way we interact with machines and make them understand our feelings. Emotional expressions can occur both with and without self-consciousness. People, though, are probably to always have conscious control over their emotional expressions.

Emotions (E=outwards, motion = movement) is a movement that goes out and plays a key role in human life as they act as motivators. These emotions can be defined as complex psychological reactions to stimulation. Emotions can be positive or negative and have a big impact on us. According to Paul Eckman [1], six basic emotions are universal: happiness, sadness, anger, fear, surprise, and disgust. However, later psychologists added to this list other emotions such as pride, excitement, embarrassment, contempt, shame, etc.

Emotion recognition and detection have shown an increased interest in various fields like Human-computer interaction [2] health monitoring [3] [4], security, mobile computing [5] [6], robotics [7], etc. Many features including facial expressions [8] [9] [10], speech [11] [12] [13], EEG, body posture [14] [15], Skin temperature Measurements (STM), and Gesture Analysis can perform emotion recognition.

Amongst all these features, Facial Expression is the most popular one due to its conveyance of people's attitudes, affects, and intentions, which explain its importance in classifying emotions. Several issues can affect the performance of an algorithm developed using computer vision techniques. For example, different subjects might express the same emotion differently. If the emotion needs to be detected from the speech, the difference between the voices of the subjects and the ambient noise can affect the final recognition performance. To recognize accurately the emotions, we are proposing an algorithm for training a classifier for emotion recognition that can leverage information from both audio and visual data.

This paper proposes a methodology to recognize emotions based on visual-audio data. Section 2 reviews the related work in the field of audio-visual emotion recognition. In section 3, the details of the proposed architecture are described followed by the experimental results and discussion in section 4. Finally, section 5 offers the conclusion.

II. RELATED WORK

The automatic Emotion Recognition (AER) multimodal proposed in [16] is based on the fusion of visible and infraRed (IR) images with speech. The authors have created a dataset called VIRI (Visible and InfraRed Images). As input, we find both image types (visible and infraRed) that will be trained using a Convolutional Neural Network (CNN) model that consists of two layers to extract the features, then the features result get combined to have a first emotion decision using a Support Vector Machine (SVM) classifier. On the other hand, an input speech will be provided by incorporating an audio spectrogram for training the Artificial Neural Network (ANN) that will develop a third CNN layer that results in the second decision. The two decisions are fed to a decision-level fusion in the second layer to get the final classification. The Ryerson

Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [17] was used for the speech pre-processing. The authors tested the model on five emotion classes (Neutral, happy, Sad, Angry, and Surprised) and different modalities (images only, infrared images only a fusion of both types of images, Speech only, and finally the fusion of visible images, infrared images, and speech). The fusion of all input types had an accuracy of 86,36%.

A hybrid deep convolutional neural network was proposed by [18] that has three representations, an audio signal representation, and two visual ones. To extract the audio features, Alex Net architecture with pre-trained weights on the ImageNet dataset [19] was applied, the audio network is composed of five convolutional layers, three max-pooling layers, and three fully connected layers. The visual network is a 2D CNN model based on the VGG-Face network, it is composed of fourteen VGG layers, ten convolutional layers, four max pooling, one flattened layer, and one fully connected layer with six neurons since there are six emotion classes: anger, disgust, fear, happiness, sadness, and surprise. The authors evaluated dimensionality reduction techniques on the global concatenated feature vector like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Different classifiers were applied to test the method's accuracy like K-Nearest Neighbors (K-NN), SVM, Logistic Regression (LR), and Gaussian Naïve Bayes (GNB). Various datasets were tested to validate the proposed method's consistency: RML, eINTERFACE05, and BAUM-1s. The network has achieved an accuracy of 82,5% on RML, 85% on eINTERFACE05, and 59,5% on BAUM-1s.

A different deep CNN architecture was proposed in [20] composed of 3 parts, the first part is 1D CNN architecture who deals with the abstract features extracted from the image sequences, the second part is 2D CNN-based architecture which extracts the features from spectrograms that had used the Short Time Fourier Transform (STFT) and the final perform the emotion recognition part composed of two fully connected (FC) layers. The authors have implemented the CREMA database for their proposed method. The first experiment tested only the video, and the second combined video and audio input. The audiovisual had the best accuracy 69,42%.

III. Proposed Method

Our core idea is that for a task like emotion classification, there should be a common latent space into which we can project data from both modalities (video and audio).

Ideally, they should be able to use information from both domains to make better classifications, rather than just using specific patterns as shown in our global proposed architecture in Fig.2.

A. Dataset

We settled on the dataset RAVDESS [17], which contains recordings of 24 professional actors vocalizing vocabulary-matching sentences with a neutral North American accent. Different emotions include expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, namely eight categories. Each expression is produced at two levels of emotional intensity (normal, and strong). We consider this dataset because it contains both faces and voices. 7536 records contain both facial images and associated sounds, so we have data in both audio and video for a particular event, and each record is tagged with a specific emotion.to summarize, our dataset contains three modality formats:

- Only audio format (16bit, 48kHz) that are separated as speech file (60trials*24actors= 1440files) and song file (44trails*23actors=1012)
- Audio-visual and video-only format (720p H.264, 48kHz) separated as speech files (60trials*24actors*2formats= 2880) and song files (44trails*23actors*2formats=2024)

B. Pre-processing of the data

We preprocess the data as follows, for audio, we take the speech data, then sample each of these recordings using Kaiser Best sampling and get the first 40 MFCCs for each recording.

As for the images, we took videos, capturing two frames every half second, each video has about 6 frames. We connect these two frames horizontally, as shown in Fig.1. We do this because we have found experimentally that adding this redundancy helps the network easily pick out important features from images.

Fig. 1. Image Frame Example

Since the number of images is now higher than the number of recordings, we simply duplicate the recordings to ensure a one-to-one mapping between our input audio and video datasets fusion architecture.

C. Proposed Architecture

We build a model shown in Fig.3 that consists of two CNN networks: an "audio" CNN that classifies emotions on the audio features, and an "image" CNN that classifies the image features.

The "audio" CNN is composed of two CNN layers with a max pool layer, whose final output is flattened to get the joint latent representation of the audio features. In addition, the "Image" CNN is also a two-layer CNN with two max pool layers whose output is also flattened to get the representation of image features. We add a small MLP (Multi-layer Perceptron)

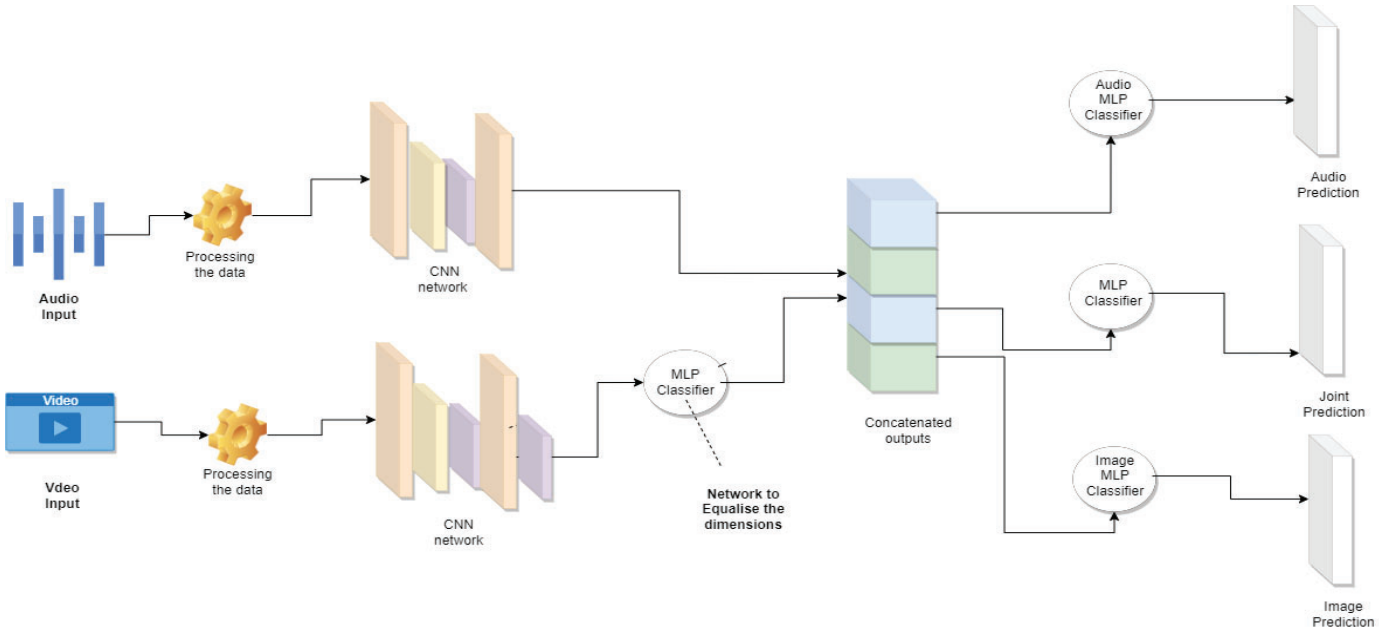

Fig. 2. Our proposed Framework for audio-visual emotion recognition

network so that Audio CNNs and video CNNs' output have the same dimension. The MLP classifier is a class of fake neural networks that oversaw learning methods returned to inciting for planning. We tried to alleviate the effect of one modality dominating the other one. Such a phenomenon can happen since the output of the image network has a higher dimension than the output of the audio part, which can lead to the image modality dominating and ignoring the contribution of the audio modality. Now, we concatenate the output of both the "Audio" and "Image" CNNs, and we add two other MLP networks that essentially play the regularizer role.

This proposed network is essentially an MLP classifier that classifies both audio and video data separately. The input of the audio MLP network is the output of the audio CNN, and the input of the image MLP network is the output of the image CNN after the dimension regularizer. Both networks also try to classify the emotion based on their respective input and have a cross-entropy loss. The loss for this proposed architecture is defined as the sum of the loss of the joint classifier and the sum of losses of both other networks.

Where each loss is a cross-entropy loss. The basic intuition behind this kind of architecture was that during training the whole architecture might favor one modality over the other and tries to learn based only on one modality (which is typically the one that gives the best accuracy individually),

to prevent that we also added these two networks which try to ensure that the representation works individually while learning the joint representation.

While providing input to this data, we make sure that both audio and video features that are being fed to the architecture belong to the same class, i.e.; if audio being fed to the network belongs to the class "sadness", then the corresponding image being fed is also belongs to the class "sadness".

The elemental impression behind such architecture was that the separate CNNs would try to master the mappings into the common inactive space for both audio and image. After we project both audio and video into the common latent space, since a video and audio should hopefully contain distinct cues and data about the emotion, joining them will aid the combined classifier categorizes the emotion better than just giving the classifier the data from only one modality.

IV. Experimental Results

This experiment aims to see how the concatenated architecture is compared against their individual counterparts.

We took the data described above and an 80-10-10 split, with 80% for training, 10% for testing, and 10% for validation.

We trained the individual CNN classifier for both "Audio" and "Image" data separately, and then we trained the proposed concatenated architecture and compared their best test accuracies with their individual counterparts.

All the networks were trained using SGD optimizer with momentum and for 30 epochs. We consider the availability of all eight classes in the dataset for this classification.

The plot of validation accuracy shown in Fig.4 for our model gives insight into how the network behaves during training as observed the fusion classifier always remains above the individual classifier which is a desirable behavior but also notices that all three classifiers saturate almost around similar epochs which further strengthens our argument that the MLP networks do not allow one modality to dominate because otherwise, one of the modalities (typically image in this case) would converge faster and then audio (if converges) would converge at a very later stage and typically at a point worse than its individual accuracy which is not the case in our architecture.

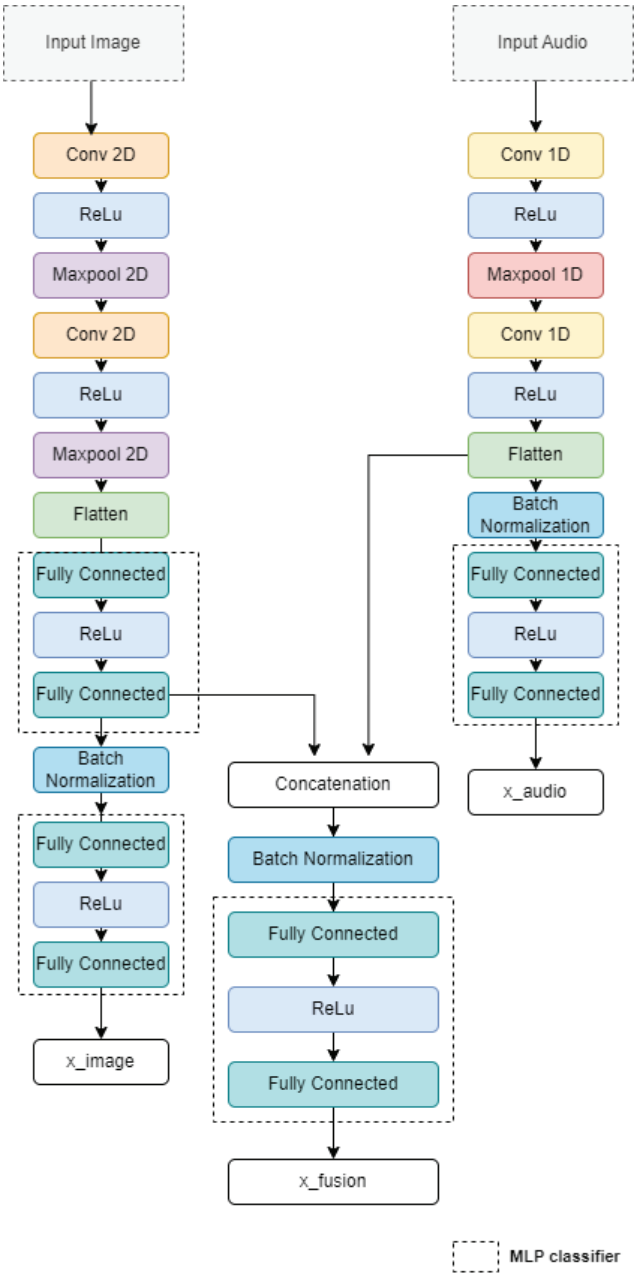

Fig. 3. Proposed Model Architecture

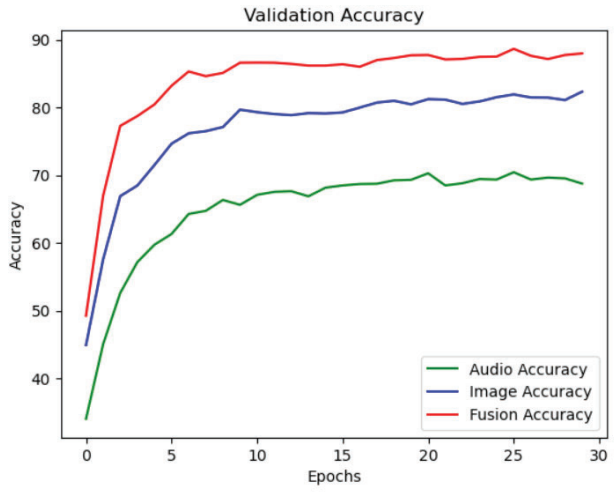

Fig. 4. Validation Accuracy for our model on different modalities

The experiments were implemented in PyTorch on a computer with an NVIDIA Quadro P2000 graphic board and Intel i7 8th GEN processor.

The obtained results are reported in Table I for different modalities for 30 epochs and essentially confirm our belief that combining data from both modalities are helpful and that projecting them into a common latent space is a feasible way to do this.

TABLE I. COMPARISON OF THE PROPOSED MODEL ON DIFFERENT MODALITIES

Network Type	Test Accuracy	Validation Accuracy	Train Accuracy
Audio	70%	70%	98%
Video	85%	85%	98%
Video+Audio	89%	89%	98%

By studying Table II, it can be observed that the proposed audiovisual framework has higher accuracy compared to some similar works done in these fields with the same dataset.

TABLE II COMPARAISON BETWEEN THE ACCURACY OF THE PROPOSED METHOD AND REPORTED RESULTS IN RECENT WORKS ON THE RAVDESS DATABASE

Method	Accuracy
[16]	86,36%.
[21]	80.08%
[22]	81,04%
Ours	**89%**

V. CONCLUSION

This paper presents a network architecture that can recognize emotion by combining visuals with audio information. The network output is a probability distribution for each type of emotion considered for training.

The results sustain the idea that having further data about a subject is substantial for determining the emotion with increa sed accuracy. In this sense, it can be observed that combining the data in the network, yields an enhancement of nearly 19% for the audio and an improvement of 4% for the image network.

In this paper, we studied how recognition of emotions is necessary to understand the user. We also saw how emotions can be expressed in more than one way. However, emotion detection still has many aspects to ameliorate in the future.

There's a proverb that says:" Strength comes from the union "which sounds perfect in emotional recognition. These hybrid models do combine both speech and facial sources to estimate a result. It is proven that combining information from different channels improves accuracy significantly. It's important to mention that the multimodal approach is not just a system that

takes the information from the face and the voice and calculates the average of each result, but should create a model that trains the network on both modalities at the same time.

As our architecture seems to have good results, we had the idea to create a real-time emotion detection tool based on this model and exploit it in other fields like human-computer interaction, psychological emotion disorder experiments, and User Experience (UX) fields.

REFERENCES

[1] P. Ekman, "An argument for basic emotions," *Cognition and emotions,* vol. 6, pp. 169-200, 1990.

[2] J. Galindo, *UI adaptation by using user emotions,* 2019.

[3] J. Torous, R. Friedman et a. M. Keshavan, «Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions,» *JMIR mHealth and uHealth,* vol. 2, n° %11, 2014.

[4] M. S. Hossain et a. G. Muhammad, «Cloud-assisted speech and face recognition framework for health monitoring,» *Mobile Networks and Applications ,* vol. 20, n° %13, p. 391–399, 2015.

[5] S. F. L. F. a. H. L. Z. Lv, «Extending touch-less interaction onvision-basedd wearable device,,» *IEEE Virtual Reality,* p. 231–232, 2015.

[6] E. C. J. D. a. B. S. Z. Zhang, «Cooperative learning and its application to emotion recognition from speech,» *Transactions on Audio, Speech, and Language Processing,* vol. 23, n° %11, p. 115–126, 2015.

[7] A. S. J. C. D. R. a. A. B. E. Russell, «Smile: A portable humanoid robot emotion interface,» chez *9th ACM/IEEE International Conference on Human-Robot Interaction, Workshop on Applications for Emotional Robots,*, Germany, 2014.

[8] D. S. A. Hussain and A. S. A. A. Balushi, "A real-time face emotion classification and recognition using deep learning model," in *International Conference on Emerging Electrical Energy, Electronics and Computing Technologies* , Malaysia, 2019.

[9] S. Minaee, M. Minael et A. Abdolrashidi, «Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network,» *Sensors,* 2021.

[10] G. P. Kusum and J. A. P. L. , "Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16," *Advances in Science, Technology and Engineering Systems Journal,* vol. 5, no. 6, pp. 315-322, 2020.

[11] C. Jin, A. Sherstneva and I. Botygin, "Speech Emotion Recognition ased on Deep Residual Convolutional Neural Network," *Eurasian scientific journal,* 2022.

[12] I. Popovic, D. Culibrk, M. Mirkovic and S. Vukmiroovic, "Automatic Speech Recognition and Natural Language Understanding for Emotion Detection in Multi-party Conversations," *MuCAI ?20: Proceedings of the 1st International Workshop on Multimodal Conversational AI,* pp. 31-38, October 2020.

[13] Z. Tariq, S. K. Shah and Y. Lee, "Speech Emotion Detection using IoT based Deep Learning for Health Care," in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA,, 2019.

[14] L. D. Lopez, P. J. Reschke, J. M. Knothe et a. E. A. Walle*, «Postural Communication of Emotion: Perception of Distinct Poses of Five Discrete Emotions,» *Frontiers in psychology,* vol. 8, n° %1710, 2017.

[15] N. A. Martin-Key, E. W. Graf, W. J. Adams et G. Fairchild, «Investigating Emotional Body Posture Recognition in Adolescents with Conduct Disorder Using Eye-Tracking Methods,» *Res Child Adolesc Psychopathol.,* vol. 7, n° %149, pp. 849-860, 2021.

[16] M. F. H. Siddiqui and A. Y.Javaid, "A Multimodal Facial Emotion Recognition Framework through the fusion of speech with visible and infrared Images," *Multimodel Technologies and Interaction,* vol. 4, no. 3, p. 46, 2020.

[17] S. R. Livingstone et F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,» *PLoS ONE,* vol. 13, n° %15, 2018.

[18] J. Y. R. Cornejo and H. Pedrini, "Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Bari, Italy., 2019.

[19] J. a. D. W. a. S. R. a. L. L.-J. a. K. L. a. L. F.-F. Deng, «ImageNet: A large-scale hierarchical image database,» chez *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[20] L.-C. D. A. R. Nicolae-Catalin Ristea, «Emotion Recognition System from Speech and Visual Information based on Convolutional Neural Networks,» *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD),* pp. 1-6, 2019.

[21] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. Montero et F. Fernández-Martínez, «Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning,» *Sensors ,* 2021.

[22] K. Aghajani, «Audio-visual emotion recognition based on a deep convolutional neural,» *Journal of Artificial Intelligence and Data Mining (JAIDM),* vol. 10, n° %14, pp. 529-537, 2022.

[23] A. M. a. B. H. a. M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing,* vol. 10, no. 1, pp. 18-31, 2019.

[24] N. A. M. L. Selma Medjden, "Adaptive user interface design and analysis using emotion recognition through facial expressions and body posture from an RGB-D sensor," *PLOS ONE 15,* vol. 15, no. 7, 2020.