IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

Multimodal Spatiotemporal Representation for Automatic Depression Level Detection

Dubale Swapnil Arun[1], Shelke Mayur Pandit[2], Kapadane Vaibhav Shyam[3], Savale Vishal Ashok[4], Prof. C. S. Wagh[5]
Students, Department of Computer Engineering[1,2,3,4]
Guide, Department of Computer Engineering[5]
Navsahyadri Education Society Group of Institutions, Naigoan, Maharashtra, India
swapnildubale04@gmail.com[1], shelkemayur4700@gmail.com[2],
amvaibhav17@gmail.com[3], vishalsavale33@gmail.com[4]

Abstract: *Humans being's cognitive system can be simulated with artificial intelligence systems. Machines and robots equipped with cognitive ability can automatically recognize a human's mental state through their gestures and facial expressions. An artificial intelligent system is proposed in this paper to monitor depression. In this world there are lot of people suffering from depression, and this impacting the economy very badly. Getting it checked and treated as soon as possible is important to cut costs and even lives can be saved too by early diagnosis of depression. It has been tried many times to figure out how depressed someone is using audio and visual features and the text analysis of conversational speech transcriptions. But it is hard to get the data and there are not lot of options for research. It is important to choose the right training data because there are not too many features that can be extracted from speech to detect depression level. Therefore the method described in this paper uses audio as well as video features to figure out how depressed someone is. This help in recognizable improvement in accuracy of detection of depression.*

Keywords: Depression, Face Expression, Audio Features, Machine Learning, Multimodal Depression detection.

I. INTRODUCTION

Depression is a psychiatric disorder that can cause people to have very low mood and they are not able to go about their normal lives. This psychiatric disorder is more serious as it can lead people to self-harm or even suicide behaviors sometimes. Mental health issues such as depression are linked to deficits of cognitive control. In 2017 World Health Organization stated that there are 350 million people are depressed worldwide and by 2030 depression will be the second most common cause of death. Proper treatment and early diagnosis can help to get out this psychiatric disorder as soon as possible. As the diagnostic process relies totally on the doctor's clinical experience some people don't get the needed treatment in time. Thus it will be helpful to develop a system which can make doctors more efficient. It has been observed that people having depression has different speech and facial movement than the healthy people not having depression. Therefore using speech and facial expressions as biomarkers to detect the depression level of peoples. Mathematical modeling and Artificial Intelligence techniques are being effectively involved in mental health research to try and solve this type of problems. The face can effectively communicate emotions among other peoples with the use of the facial expressions. However, Depression disorders are not limited to be expressed by face. The perception of emotional body movements and gestures has shown it can be observed through a series of controlled experiments using patients with and without depression.

II. LITERATURE SURVEY

It can be observed that recent years witnessed an increase of research for clinical and mental health analysis with the help of facial and vocal expressions. Emotion recognition from facial expressions has progressed significantly. Proposed system help to automatically quantify emotional expression differences between patients with psychiatric disorders and healthy individuals. Making man-machine interface more flexible and easy to use for the user is the main motivation of this research. Human experts will have privileged knowledge that codes the facial features of aging, such as smoothness, face structure,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

skin inflammation, lines under eye bags, when determining aged. Privileged knowledge is unavailable for test images in automated age estimates. We hypothesize that asymmetric data may be used to resolve this problem. To enhance the generalization of the trained model, be explored and exploited.

III. Proposed System

There have been lots of studies regarding Depression and mental health. What stands out the most about these studies is the conclusion that depressed person and healthy person have significant difference when it comes to their voice texture and facial expressions. Our subconscious mind lets our emotions come forth involuntarily. That's why this system is so effective when it comes to Depression prediction. To achieve our target, we divide all the audio spectrums and video segments into fixed length and input it into the STA. This STA after processing the data return ASLF and VSLF which we use to tell whether the person is depressed or not. EEP method is there to give us the aggregated ASLF and VSLF which gives us ALF and VLF.

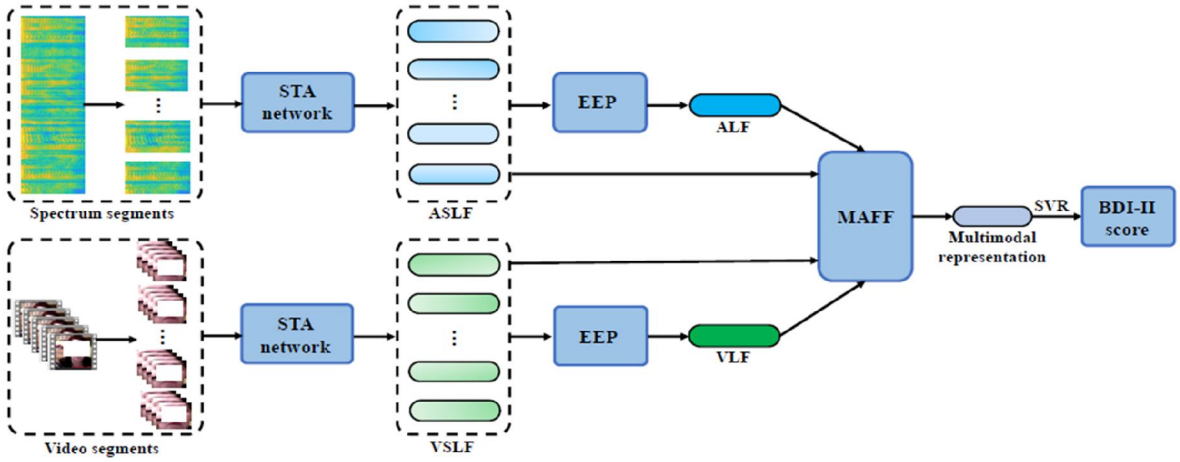


Fig. 2. The illustration of the multimodal spatiotemporal representation framework for automatic depression level detection. Our method firstly inputs spectrum/video segments into the STA network and takes the output of the last full connection layer as the ASLF or VSLF. Then, we use the EEP method to aggregate these ASLFs or VSLFs into the ALF or VLF. Finally, the multimodal representation is generated using the MAFF strategy and input into the SVR to predict the BDI-II score.

After that we Fuse the ALF and VLF obtained together using MAFF strategy which is Multimodal Attention Feature Fusion. The Fusion of these ALF and VLF gives us the BDI score. This BDI score is used to determine the level of depression a person is suffering from.

IV. Implementation

We have used dual modality in the sense there are more than one mode through which we are going to determine the depression of an individual. To do that we need VSLF and ASLF to give us individual output of each mode. We use STA network to extract ASLF and VSLF. In ASLF we use 2d CNN to segmentize the audio file. Which then is converted into 1d array with the help of flatten. LSTM is used to extract temporal sequence. Fully connected layers feed forward the output of previous data sequence to next function. We can represent it mathematically as follows,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

$$\mathrm{w_A} = \mathrm{softmax}\left(\hat{\mathrm{w}}_\mathrm{A}\right) = \frac{\left[\exp\left(\hat{w}_1\right), \cdots, \exp\left(\hat{w}_{T_\mathrm{A}}\right)\right]^\mathsf{T}}{\sum_{t_\mathrm{A}=1}^{T_\mathrm{A}} \exp\left(\hat{w}_{t_\mathrm{A}}\right)}, \quad (1)$$

where "T" refers to matrix transposition and $\hat{\mathrm{w}}_\mathrm{A} \in \mathbb{R}^{T_\mathrm{A} \times 1}$ is calculated by Eq. (2).

$$\hat{\mathrm{w}}_\mathrm{A} = \mathrm{o}_\mathrm{A}^{\mathrm{L}\,\mathsf{T}} \odot \mathrm{o}_\mathrm{A}^\mathrm{C} \triangleq \left[\hat{w}_1, \cdots, \hat{w}_{T_\mathrm{A}}\right]^\mathsf{T}, \quad (2)$$

where "$\odot$" represents the matrix multiplication operation, which captures the correlation between $\mathrm{o}_\mathrm{A}^\mathrm{C}$ and each frame feature of the $\mathrm{o}_\mathrm{A}^\mathrm{L}$ for integrating the spatiotemporal information and generating the spatiotemporal attention weight.

$$\mathrm{r_A} = \mathrm{o}_\mathrm{A}^\mathrm{L} \odot \mathrm{w_A}. \quad (3)$$

We need 3d CNN to segmentize the Video features and get another vector. These features again using Flatten are converted into 1d array. Feed forward circuits do their job and keep forwarding the outputs which are then multiplied in the form of matrices. SoftMax converts these results into 1 and 0 which helps us predict with ease.

$$\mathbf{G}^* = \underset{\mathbf{G}^\mathsf{T}\mathbf{G}=\mathbf{I}_K}{\arg\min} \; -\sum_{i=1}^{D}\sum_{k=1}^{K}\left(\mathbf{g}_k^\mathsf{T}\mathbf{d}_i^\mathsf{T}\mathbf{g}_k\mathbf{d}_i\right). \quad (5)$$

$$\mathbf{G}^* = \underset{\mathbf{G}^\mathsf{T}\mathbf{G}=\mathbf{I}_K}{\arg\max} \; \sum_{k=1}^{K}\mathbf{g}_k^\mathsf{T}(\mathbf{S}^\mathsf{T}\mathbf{S})\mathbf{g}_k. \quad (6)$$

$$\mathbf{S}^\mathsf{T}\mathbf{S} = \sum_{m=1}^{M}\lambda_m \mathbf{q}_m \mathbf{q}_m^\mathsf{T}, \; \lambda_1 \geq \cdots \lambda_M, \quad (7)$$

where $\lambda_m \; (m = 1, \cdots, M)$ is the m-th eigenvalue of $\mathbf{S}^\mathsf{T}\mathbf{S}$ and $\mathbf{q}_m \in \mathbb{R}^{M \times 1}$ is the corresponding eigenvector. Note that $\mathbf{q}_m{}^\mathsf{T}\mathbf{q}_m = 1$ and $\mathbf{q}_r{}^\mathsf{T}\mathbf{q}_t = 0, \; r \neq t$.

$$\mathbf{G}^* = \underset{\mathbf{G}^\mathsf{T}\mathbf{G}=\mathbf{I}_K}{\arg\max} \sum_{k=1}^{K}\sum_{m=1}^{M}\lambda_m\left(\mathbf{q}_m^\mathsf{T}\mathbf{g}_k\right)^2. \quad (8)$$

$$\mathbf{G}^* \leq \underset{\mathbf{G}^\mathsf{T}\mathbf{G}=\mathbf{I}_K}{\arg\max} M \sum_{k=1}^{K}\sum_{m=1}^{M}\lambda_m\left[\sum_{i=1}^{M}\left(q_{m_i}g_{k_i}\right)^2\right], \quad (9)$$

Now these mathematically solved equations make our stand clear and give us the BDI scores.

V. Non Functional Requirements

5.1 Performance Requirements

The performance of the functions and every module must be good. The overall performance of the software will permit the users to work efficiently. Performance of encryption of data must be fast. Performance of the providing virtual environment should be a fast Safety Requirement. To detect and fix the errors the application is designed inappropriate modules. This makes it easier to install and update new functionality if any is required.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

5.2 Safety Requirement

The application is invented in modules where errors can be detected and fixed easily. This makes it easier to install and update new functionality if any is required.

Our software is having many quality attribute that are given below

- **Adaptability:** This software is adaptable by all the users.
- **Availability:** This software is freely available for all the users. The availability of the software is easy for anyone to use**.**
- **Maintainability:** After the deployment of the project if any error arises then it can be easily maintained by the software developer.
- **Reliability:** The performance of the software is best which will increase the reliability of the Software.
- **User Friendliness:** Since, the software is a GUI application; the output generated is much user friendly in its behavior and it's very easy to operate.
- **Security:** Users are authenticated using security phases so reliable security is provided.
- **Testability:** The software will be tested considering all the possible aspects

VI. Experimental Results

A. Depression Detection Performance using Different Modalities and Network Structures

Based on the above setup, we examine the detection performance using different modalities and network structures through predicting the depression level on the development sets of AVEC2013 and AVEC2014. The CCL means to concatenate the outputs of 2D CNN (3D CNN) and LSTM (2D CNN+LSTM) without using the attention mechanism, where the output of LSTM (2D CNN+LSTM) is a vector rather than a sequence. Note that the EEP and SVR are for aggregation and prediction in these experiments. For the detection accuracy using the audio modality, the LSTM performs better than the 2D CNN, which is due to the fact that the spectrum is a sequential data and temporal changes are more expressive than the spatial structure in characterizing the depression cues. The reason for good performance of the CCL and STA network is that they both contain the spatial and temporal information. The result also illustrates that the spatial and temporal features of the spectrum are both helpful for depression detection. Moreover, the STA network performs better than the CCL. The reason is that the STA network uses attention mechanisms to integrate spatiotemporal information and emphasize frames related to depression. And CCL is limited in selecting key frames.

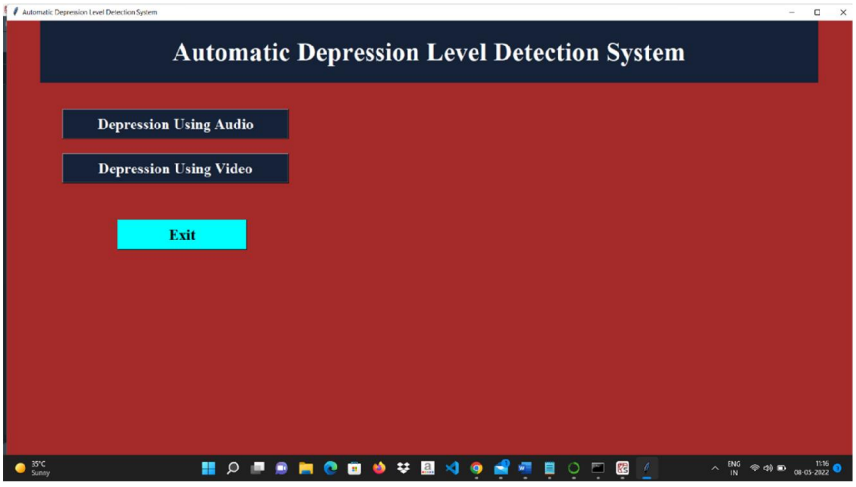
B. Depression Detection Performance using Different Information

In this section, we validate the detection performance of different information on the development sets of AVEC2013 and AVEC2014 databases. Note that the STA network, EEP and SVR are for the segment-level feature extraction, aggregation and prediction. One can see that the detection performance of VAAF is better than that of ALF. The reason is that the more pronounced depression cues are contained in facial activity than speech. Thus, the usage of attention mechanisms between VLF and ASLFs can extract the information similar to video from audio and improve the prediction accuracy. The similar reason can be explained by the result that the AAVF is not as good as VLF. In addition, we find that “ALF+VAAF” and “VLF+AAVF” both obtain better detection performance than “ALF” and “VLF”. This result illustrates that the modal complementary information (i.e., VAAF and AAVF) is helpful to improve the experimental accuracy. “VCA + VAAF + AAVF” achieves the best detection result because it contains not only audio and video features, but also their complementary information.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

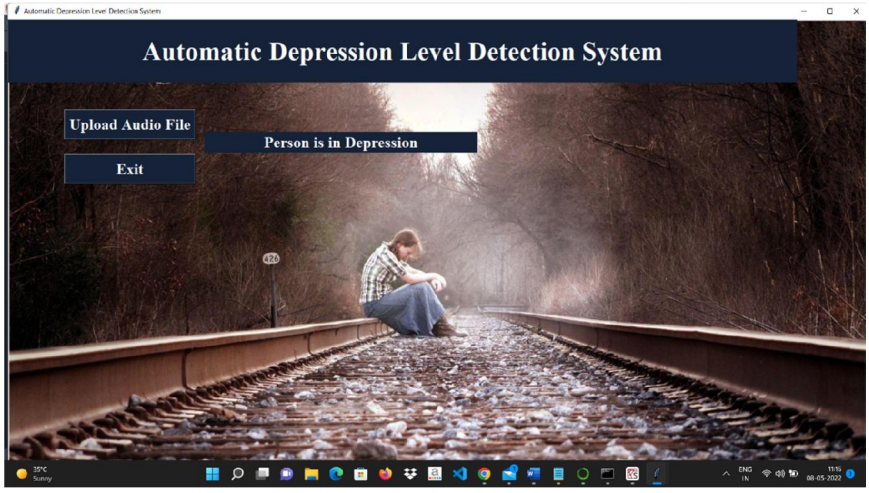
VII. RESULTS

The main GUI of our Project looks like below Image



After calculating Depression by using the video result looks like below

After calculating Depression by using the Audio result looks like below



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

VIII. Conclusion

Depression Detection System is designed to look for signs of depression in people's images that have been taken by people who have been approved by the system. This makes the system safe. The picture is taken automatically when the user is logged in at a certain time. The images that are taken are used to figure out how depressed the user is based on some standard conversion and image processing methods. Then the system will use Machine Learning algorithms to see if the person is depressed or not. This makes the results more efficient.

References

[1]. Niu M, Tao J, Liu B, et al. Automatic Depression Level Detection via lp-norm Pooling[C]//2019 Conference of the International Speech Communication Association (INTERSPEECH 2019). 2019: 4559-4563

[2]. Zhou X, Jin K, Shang Y, et al. Visually Interpretable Representation Learning for Depression Recognition from Facial Images[J]. IEEE Transactions on Affective Computing, 2018.

[3]. Lee J, Kim S, Kiim S, et al. Spatiotemporal Attention Based Deep Neural Networks for Emotion Recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 1513-1517.

[4]. De Melo W C, Granger E, Hadid A, et al. Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions[C]//2019 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019: 1-8

[5]. Al Jazaery M, Guo G. Video-based Depression Level Analysis by Encoding Deep Spatiotemporal Features[J]. IEEE Transactions on Affective Computing, 2018.