IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Life Logging Using Egocentric Perception

A. Anandakrishnan[1*] , A. Walia[2] , A. Jha[3], J. J. Pandya[4], C. V. Raj[5]

[1,2,3,4,5]Dept. of Computer Science and Engineering, The National Institute Of Engineering, Mananthavadi Road, Mysuru, 570008, India

Corresponding Author: akshyan97@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si14.335338 | Available online at: www.ijcseonline.org

Abstract:-This paper aims to provide a solution that uses a multimodal approach to analyse large intake of audio and video data and use it to understand the emotions of a subject and to describe the current surroundings to the subject in question. The model is trained on the egocentric data, which contains audio and video signals. The model contains emotion recognition and a speech recognition which extract features of their own allowing to perform a classification on the emotions. The large inflow of data from up and coming technologies like Google Lens and onset of Internet of things are key application points for this solution.

Keywords:-Emotion Recognition, Face Extraction, Speech Recognition,Scene Description,Life Logging

I. INTRODUCTION

There are different modes of communication possible for human being like speech, facial expressions or emotions. The fastest among these is speech which can be extracted in the form of signals and can be studied upon to yield the pattern and intentions of the speaker. Similar studies can be performed on video of a subject which provides important data points like facial expression. The above mentioned information regarding speech patterns and emotions can be used wisely to improve communication between humans and machines which in turn opens a wide area for technological research.

The idea of this research is to use the emotion and speech giveaway patterns and develop a more educated information base regarding a human subject as well as his surroundings. Hence the title "Life Logging".

To achieve this goal two key divisions are utilised:
Speech Recognition (AVSR): A multi-modal which uses audio and video to determine what the person is talking. Alternatively,called AVSR.
This is achieved by using a Bidirectional GRU, Convolutional Network and a Residual Network.

Emotion Recognition: The sentiment of the person describes his speech better.Adding the emotion recognition modal built using a bidirectional LSTM, Residual Network and

Convolutional Network would help us describe the sentiment of the person.

The combination of above two can be used for various real life applications such as lie detection, scene description, emotion detection.This is achieved using training based on two main datasets one is Vidtimit and another is RECOLA dataset.

II. DATASETS

The two main inputs required for training of the desired modals are the large datasets on which pattern recognition can be performed iteratively. The datasets used for the solution are as follows:

Vidtimit Dataset: This dataset is a collection of data containing the speeches and the audio of 48 people. The sentences recited by the individual are different except for the first two. This dataset is mainly used for the training of our speech recognition model which runs a bidirectional gated recurrent network.

This dataset helped us extract features such as the head rotation and the movement of the lips for the lip reading and the most important one, the wavelets for the audio which is passed through a residual network for the feature extraction. This dataset gave us a word error rate of 15.6% and the sentence error rate of 13.6%. This helps in recognizing and converting speech to text in an accurate and more efficient way.

Recola Dataset: This database or the dataset was developed by the researchers at the University of Switzerland, which is a multi model dataset, meaning the dataset can be trained on multiple models and can extract multiple features for a given

scenario. This database has mainly the different gestures and the facial emotions given by 46 French people. This dataset contains audio, videos from which our residual and 3D conv network extracted features to classify the different emotions. For extraction and learning of more features these videos and audios were passed through long short-term memory, a recurrent network to get the classification of emotions. The dataset gave us an accuracy for happy of 80% and the accuracy for anger around 67%. This can be improved by training it on a larger dataset.

III. LITERATURE SURVEY

The Audio-Visual Speech Recognition: There has been many works related to the speech recognition which is based out of the only the audio signals which are transformed into waves and processed through a deep neural network where a lot of work has been done and will not be discussed here. Our main objective is the use of the lip reading technique which involves the video of the person talking. A non neural network method by Abiel Gutierrez et al[1] which would be tedious and slow and low level of accuracy could be done using the computer vision by studying the texture and analysing the frames of the video. The lip reading can be done using the Convolutional Neural Network which was one of the most used ways. TriantafyllosAfouras et al [2] uses the WLAS model and the transformer model to perform the combination of the audio and the lip reading video signals to extract features which later is then basically sent through a classification layer. Satoshi Tamura et al [3] studied and extracted the bottleneck features[4].

The Emotion Recognition, The emotion recognition has been in field of artificial intelligence for quite some time now. A lot of different ways has been setup. Emotions can be studied by multiple things like gestures, speech and the most importantly the facial expressions. There has been a lot of work which involved the use of the convolutional neural network which basically brought the use of the region of interest which has be studied by M Xu et al[5] . The emotions can be recognized using the speech that has been extracted from the speech recognition using the concept of sentiment analysis.HPalang et al [6] performed a sentiment analysis on the sentences using sentence embedding and long short term memory(LSTM) which had gained a huge amount of fame for its accuracy. But this later was replaced by the gated recurrent network built by Wang X et al[7]. Finally for the emotion recognition with the help of a multi-model can be built which is going to be our main focus. This area hasn't been researched as much but still there has been some work with an accuracy of only 50-60%. P. Tzirakis et al [8] worked on an end to end system which uses an autoencoder for the emotion recognition and uses all CNN to extract features from the image like gestures, expressions and so on. The same thing is being utilised with a dataset called the recola and improving it with a LSTM based structure as shown later.

IV. WORKING

The life logging happens mainly in two important steps which was discussed separately in the introduction. The speech recognition and the emotion recognition. The entire model or the flow of the project is shown in figure1.

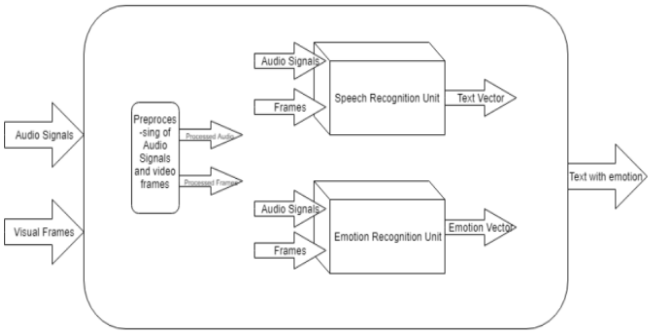

Figure 1: The overall architecture

The figure shows the basic flow of the entire system. The first thing performedwasseparating the audio and video(frames) into two different signals. This is done in the pre-processing part of the system. Before passing the frames which are normalized and the audio signals are transformed to wave images.

The Speech Recognition:- The multi-model speech recognition which extracts features from the audio waves using the technology used by the wavenet and the soundnet is being used. This vector along with the features extracted from the lip reading is sent to the classification layer giving us a text vector. The bidirectional gated recurrent network is used in the learning of the words depicted and remembering the words which then helps us classify words based on the percentage of the words database. The workflow of the speech recognition can be seen in the figure 2

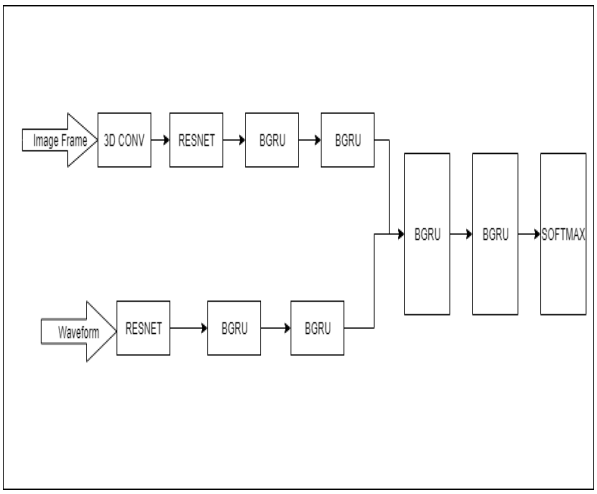

Figure 2 : The speech recognition model

The Emotion Recognition :- The emotion recognition is the part where the emotions are determined using the speech vector and the gestures and the other features that have been extracted can be classified into different emotions shown by a person. This can be done using a bidirectional lstm in an end to end system. This is depicted in the figure 1.3 .

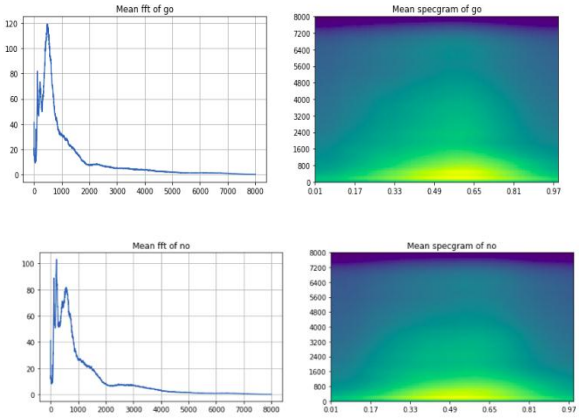

Figure 3: The graph of two words (no and go)

The bidirectional lstm is only for the recognition of emotion. The feature extraction is mainly done in the resnet and the convolutional network. The preprocessing is the same as speech recognition. The emotion decoding problem can be resolved using : $l^* = \arg\max_l AM(l) + \alpha LM(B(l))$
We are also performing a sentiment analysis on the text extracted which becomes one of the features.

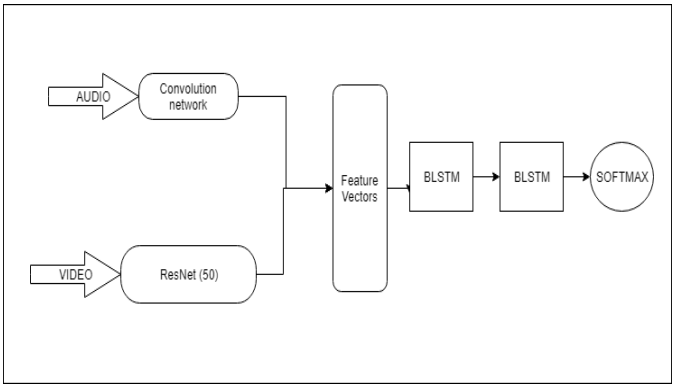

Figure 4: The emotion recognition model

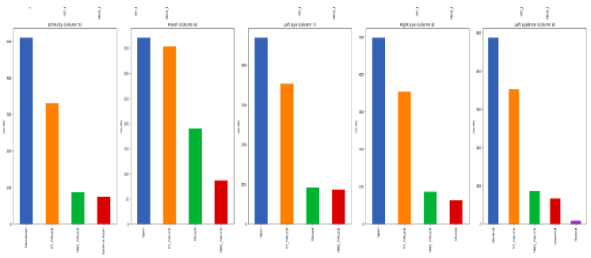

Figure 5: The recola dataset graph representation of 4 emotions on 4 people

V. RESULT

The speech recognition was trained on the vidtimit dataset. Based on different number of epochs the different word error rate can be determined as shown in the table-1

Table-1 Speech Recognition Outputs	Weights(M)	Epochs	Word-Error Rate(%)
	3.8	150	10.6
	3.8	75	15
	4.3	112	12
	4.3	144	7.8
	5.6	134	11
	5.6	144	8.5
	5.6	155	8.3
	6.2	120	10.3
	6.5	122	10.6

The emotion recognition which was trained on the recola database gave us the following outputs where the accuracy ranges from the 45-64 for different number of epochs (number of feed forward and back propagation). The F1 score is the number of correct against all the positive depictions.

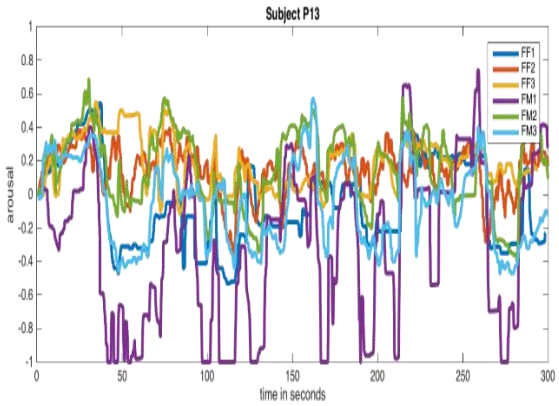

Figure 4: The activation of emotion recognition can be seen for the various emotions in the recola dataset

The emotion recognition which was trained on the recola database gave us the following outputs as shown in table-2 where the accuracy ranges from the 45-64 for different number of epochs (number of feed forward and back

propagation). The F1 score is the number of correct against all the positive depictions.

Table-2: Emotion Recognition Observations	Epochs	Happy		Happy	
		Accuracy	F1	Accuracy	F1
	70	45	43.2	34	35.2
	140	53	52.4	45	42.6
	180	60.4	63.3	57.5	56,7
	240	56.4	54.3	64.3	63.2
	260	57.3	57.3	60.6	63.3
	280	60.2	58.2	58.4	55.8
	300	63.4	61.3	59.3	57.4
	320	65.5	63.4	54.5	52.4

VI.CONCLUSION AND FUTURE WORK

From the above experiments the speech and the emotions of a person can be extracted which allows to perform a life logging of an individual. This helps to keep track of the events that occur on a day to day basis. This is done using the deep neural network approach like the convolutional neural network and recurrent network, which are combined to form a multi model scenario.

This can further be improved by adding a scene description where a convolutional network along with the reinforcement algorithm can be used to depict the scenario the person is in and add it to the life logging event.

REFERENCES

[1] A. Gutierrez and Z. Robert. Lip Reading Word Classification

[2] Ashish B. Ingale,.D.S.Chaudhari "Speech Emotion Recognition" in International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[3] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu, “Audio-visual speech recognition using deep bottleneck features and high-performance lipreading,” in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015, pp. 575–582

[4] Z. Ma, H. Yu, W. Chen, and J. Guo, “Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features,” IEEE Trans. Veh. Technol., vol. 68, no. 1, pp. 1–13, 2019.

[5] M. Xu, X. Deng, S. Li and Z. Wang, "Region-of-Interest Based Conversational HEVC Coding with Hierarchical Perception Model of Face," in IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 3, pp. 475-489, June 2014.

[6] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(4):694–707, 2016.

[7] Wang X, Jiang W, Luo Z. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In Proceedings of the InternationalConference on Computational Linguistics (COLING 2016),2016

[8] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

