Mouth Gesture Interface for VLC Media Player

A.Rajaprabu, A.Geetha

Abstract— This proposed work presents a frame work of mouth gesture recognition for Human Computer Interface (HCI). It replaces the traditional input devices such as mouse and keyboard which allows a user to work on a computer using his/her mouth gestures. This work is aimed at helping severely disabled and paralyzed people. The entire work includes mouth detection, region extraction, gesture classification, and interface creation with computer applications. Initially face and mouth regions are detected using Haar-cascaded classifier. Secondly, the gesture recognition is done using the concept of Deep learning through Convolutional Neural Network (CNN). The mouth gestures are recognized and classified as mouth close, mouth open, tongue left and tongue right. Finally an HCI is created by mapping the mouth gestures into VLC player operations such as play, pause, forward jump and backward jump. The performance of the proposed method is measured and compared with other existing methods. This work is found to perform better than the other methods.

Keywords: Human computer interface, gestures, cascaded classifier, Deeplearning, Convolutional Neural Network

I. INTRODUCTION

Now a day the need of Human Computer Interaction is increased. HCI is used when the situation arises when not able to use the tradiditional input devices like keyboard and mouse. The main goal of making HCI application is to work with computer in natural, intuitive and effective ways as much as possible. The Human Computer Interface can be described as the point of communication between the user and the computer. The flow of information between the human and computer is defined as the loop of interaction.Researchers in the field of HCI both observe the ways in which humans interact with computers and design technologies that let humans interact with computers in novel ways. There are two main categories to implement such kind of problems. The first category used devices mounted on the human body directly. The second category is contactless and remote sensors. Among the contact less category, vision based human computer interfaces are the most useful one. This vision based human Computer interface uses cameras and image processing algorithms. Many vision based algorithms are implemented with eye and hand gestures. In this proposed work mouth gestures are used to create HCI with computer applications.

II. RELATED WORKS

PiotarDalka [1] presents an algorithm for lip movement tracking and lip gesture recognition for the purpose of Multimodal Human Computer Interface. It is specially used for severely disabled and paralyzed people. Lip gesture recognition is done by Artificial Neural Network and accurate lip shape segmentation is done by Fuzzy clustering.

Marcelo Archajo Jos [2] discusses the Lip control system. It's an innovative human computer interface specially designed for people with tetroplegia. This work provides a method of using lower lip potential to control an input device. It is found very much useful in developing assistive technologies.

MargritBetke [3] deals with the "Camera Mouse" system which has been developed to provide computer access for people with severe disabilities. The system tracks the user's movements with a video camera and translates them into the movements of the mouse pointer on the screen.

Yo-Jen Tu [4] presents a face and gesture based human computer interaction (HCI) system.Here combination of head pose and hand gestures are used to control the system.

MrunalineePatole [5] presents an evaluation of the lower lip potential to control an input device. A Lip mouse system is a non-prominent method that helps a user to work on a computer using movements and gestures made with mouth only, especially for handicapped people.

K.Meenakshi [6] deals with interface that allows people to move cursor and perform mouse click operations using mouth gestures without any special hardware.

III. PROPOSED WORK

In this proposed work mouth gestures based Human Computer Interface is developed that allows users to work on a computer using movements and gestures made with his or her mouth only. The main task of the proposed work is to detect and analyze images of user's mouth region in a video stream. In this proposed work four mouth gestures are detected. They are mouth open, mouth close, tongue left and tongue right. Each gesture is associated with a specific action. The mouth open and mouth close gestures are mapped with play and pause operation in VLC media player correspondingly. Likewise tongue left and tongue right gestures are mapped in to a backward jump and forward jump operations of VLC player.

Revised Manuscript Received on April 12, 2019.

A.Rajaprabu Research Scholar, Department of Computer Science & Engineering Annamalai University,Annamalainagar, Taminadu, India. (E-mail: rajaprabu_ei@yahoo.co.in)

Dr. A.Geetha Associate Professor, Department of Computer Science & Engineering Annamalai University,Annamalainagar, Taminadu, India. (E-mail: aucsegeetha@yahoo.com)

IV. DESIGN OF THE PROPOSED WORK

The flow diagram of the proposed work is shown in Fig1.The input video contains mouth close, mouth open, tongue left and tongue right gestures. From the input video face region is detected and then mouth region is detected. Haar-cascaded classifiers are used for detection. Then the CNN is trained with all the four types of mouth gestures. Finally the mouth gestures are mapped in to the operations of VLC player.

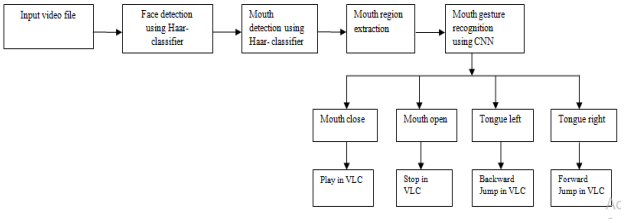

Fig 1.Flow chart for Mouth gesture recognition for Human Computer Interface

4.1 FACE AND MOUTH REGION DETECTION

A cascade of boosted classifiers working with haar like features is used to detect a user's face in the images captured by a web camera. It is an efficient algorithm for visual object detection. Face region detection shown in Fig 2 and mouth region detection shown in Fig3.

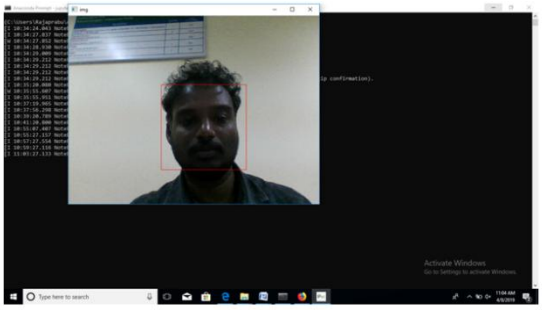

Fig2.Face detection using Haar-cascaded classifier

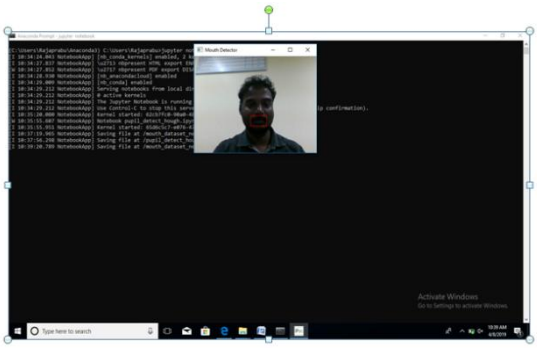

Fig3.Mouth detection using Haar-cascaded classifier

Fig4 (a) Mouth close gestures

Fig4 (b) Mouth open gestures

Fig 4(c) Tongue left gestures

Fig 4(d) Tongue right gestures4.2MOUTH GESTURE RECOGNITION WITH CONVOLUTIONAL NEURAL NETWORK

The mouth region [8] is resized into 72X43 and given as input to the model. Before passing the extracted mouth image frames in to the model it is resized in to 72X 43 pixels. The model is trained with four type gestures such as mouth close, mouth open, tongue left and tongue right gesture frames as shown in Fig 4. Here mouth close gesture labeled as 0, mouth open gesture labeled as 1, tongue left gesture labeled as 2 and tongue right gestures labeled as 3.

Mouth gestures [7] used in this proposed work are mouth close, mouth open, tongue left and tongue right as shown in Fig4(a),Fig4(b),Fig4(c) and Fig4(d). Mouth gesture recognition is done with the help of Convolutional Neural Network. Convolutional Neural Network comes under Deep learning approach. Here feature extraction task done automatically.

4.2.1 CONVOLUTIONAL NEURAL NETWORK:

The architecture of Convolutional Neural Network is shown in Fig5. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity.Convolutional Neural Network (CNN) is a special type of multilayer feed forward neural

network (MLFFNN) that is well suited for pattern classification. Development of CNN is neuro-biologically motivated.A CNN is an MLFFNN designed specifically to recognize 2-dimensional shapes with a high degree of invariance to translation, scaling, skewing and other forms of distortion.

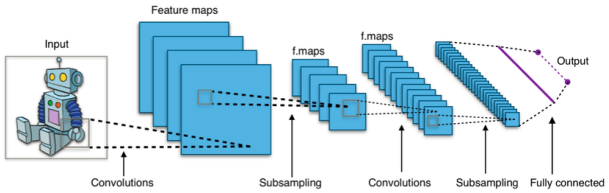

Fig5. Typical Architecture of CNN

Generally, the structure of CNN [9] includes two layers one is feature extraction layer, the input of each neuron is connected to the local receptive fields of the previous layer, and extracts the local feature. Once the local features are extracted, the positional relationship between it and other features also will be determined. The other is feature map layer, each computing layer of the network is composed of a plurality of feature map. Every feature map is a plane, the weight of the neurons in the plane are equal. The structure of feature map uses the sigmoid function as activation function of the convolution network, which makes the feature map have shift invariance. Besides, since the neurons in the same mapping plane share weight, the number of free parameters of the network is reduced. Each convolution layer in the convolution neural network is followed by a computing layer which is used to calculate the local average and the second extract, this unique two feature extraction structure reduces the resolution. Since the feature detection layer of CNN learns by training data, it avoids explicit feature extraction and implicitly learns from the training data when we use CNN. Furthermore, the neurons in the same feature map plane have the identical weight, so the network can study concurrently. This is a major advantage of the convolution network which avoids the complexity of data reconstruction in feature extraction and classification process.

4.3 TRAINING AND TESTING OF CNN

The video for input to be given to CNN [10] is captured from 40 subjects. The video tracks the face regions of the subjects and the mouth regions are extracted from the face regions using Haar-cascaded classifier. Four kinds of mouth gestures such mouth open, mouth close, tongue left, tongue right are labelled with four different classes for training the CNN. About 1000 mouth gestures containing 250 gestures of each kind are given for training.The Training Phase of CNN took 10 epochs to get trained.

For testing about 300 different mouth gestures are used and the performance of the proposed method is measured. The system is found to perform better than other testing methods.

V. INTERFACE WITH VLC PLAYER

In this proposed work, four mouth gestures are mapped into the interface operation [11] with VLC player. Mouth close gesture is mapped into pause operation and mouth open gesture [12] is mapped into play operation. Likewise tongue left and tongue right gestures are mapped into backward jump and forward jump operation. Pyautogui is the python package used for the implementation of the VLC interface.

5.1. SCREEN SHOTS

Fig.6 shows the actual screen shots of interfacing with VLC media player.

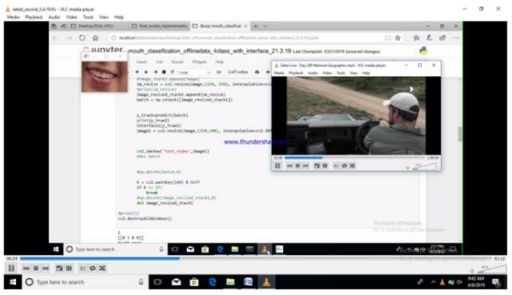

Fig 6(a).VLC interface for mouth open gesture

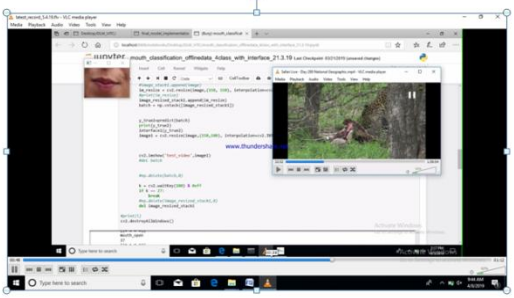

Fig 6(b).VLC interface for mouth close gesture

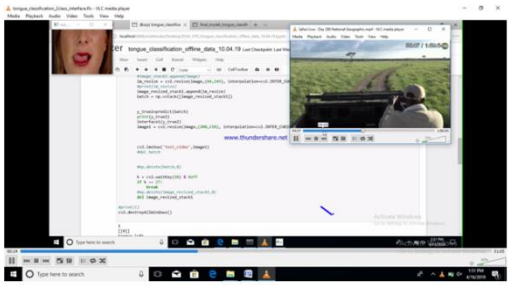

Fig 6(c).VLC interface for Tongue left gesture

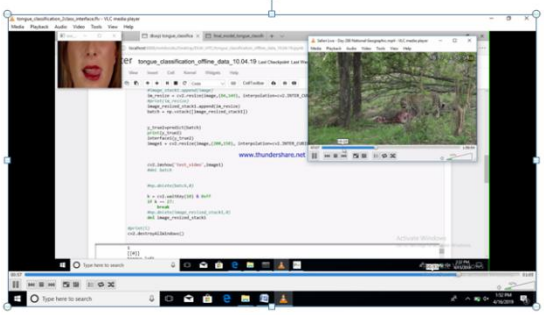

Fig 6(d).VLC interface for Tongue right gesture

VI. PERFORMANCE ANALYSIS& RESULTS

The performance of the proposed system is analyzed using the metrics as below and the results are shown.

6.1 Confusion matrix

Confusion Matrix is a summary of prediction results on a classification problem. The numbers of correct and incorrect predictions are summarized with count values broken down by each class. The Confusion matrix shows the ways in which the classification model is confused when it makes predictions.

The video containing a total of 300 mouth gestures with 95 mouth open frames, 118 mouth close frames,90 tongue left frames and 35 tongue right frames used for testing the CNN.

The confusion matrix is generated as shown in the Table 1.

Table 1: Calculated Confusion Matrix

		Predicted classes			
		Mouth open	Mouth close	Tongue left	Tongue right
Actual class	Mouth open	80	8	4	3
	Mouth close	5	108	2	3
	Tongue left	0	0	75	15
	Tongue right	0	0	5	30

6.2 Precision, Recall and Accuracy

Precision is the fraction of relevant instances among the retrieved instances.

Recall is the fraction of relevant instances that have been retrieved over the total amount of instances.

From the calculated confusion matrix, Precision, Recall and accuracy values are estimated and as shown in Table 2.

Table 2: Precision, Recall and accuracy

	Precision (%)	Recall (%)	Accuracy (%)
Mouth open	84.20	94.11	94.00
Mouth close	91.52	93.10	94.67
Tongue left	83.33	87.20	92.30
Tongue right	85.71	58.82	92.30
Avg /Total	86.19	83.30	93.31

Accuracy is the ratio of correctly predicted observation to the total observations. The proposed system is found to perform better than the existing systems found in the literature and produces an accuracy of 93.31% which is very much a satisfactory performance.

VII. CONCLUSION AND FUTURE WORK

The proposed system aims at developing a hands free HCI using mouth gestures. The system is found to produce satisfactory performance with 93.31% accuracy. The developed system is tested with VLC Media player operations and found to perform well. The system can be extended to recognize more mouth gestures and also can be applied to more interfaces for browsing, mobile applications, GPS operations etc.

VIII. REFERENCES

1. PiotarDalka,"Lip movement and gesture recognition for a multimodal human computer interface", IEEE, 2009.
2. Marcelo Archajo Jos´ and Roseli de Deus Lopes, "Human computer interface controlled by lip", IEEE Journal of Biomedical and Health Informatics, Vol.19, No.1, January 2015
3. MargritBetke, James Gips,"The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access for People With Severe Disabilities", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol.10, No.1, March 2002.
4. Yo-Jen Tu1, Chung-Chieh Kao1, Huei-Yung Lin1 and Chin-Chen Chang, "Face and Gesture Based Human Computer Interaction",International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.8, No.9, pp.219-228,2015.
5. MrunalineePatole, PoojaAthalye, "Controlling Computer with Human Lip: HCI", Journal of Software Engineering & Software Testing, Vol.2 Issue 1.
6. K.Meenakshi, Dr.A.Geetha and A.RajaPrabu, "Mouth Gestures as Human-Computer Interface", International Journal for Science and Advance Research in Technology Vol.4. No.4, pp. 503-509, April 2018.
7. GozdeYolcu, Ismail Oztel, Serap Kazan, "Deep Learning-based Facial Expression Recognition for Monitoring Neurological Disorders", IEEE International Conference on Bioinformatics and Biomedicine (BIBM),2017.
8. Luis Ricardo Sapaico, Masayuki Nakajima," Detection of Tongue Protrusion Gestures from Video", IEICE TRANS. Inf. &Syst, Vol.E94–D, No.8 August 2011.
9. Heechul Jung1, Sihaeng Lee, Sunjeong Park, Injae Lee, ChunghyunAhn "Development of Deep Learning-based Facial Expression Recognition System", MSIP(Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2014.
10. E. Perini, S. Soria, A. Prati, and R. Cucchiara "Face Mouse:A Human-Computer Interface for Tetraplegic People", Springer HCI/ECCV 2006, LNCS 3979, pp. 99 – 108, 2006.
11. Michael J. Lyons, **"**Facial Gesture Interfaces for Expression and Communication", Proceedings, IEEE International Conference on Systems, Man, Cybernetics, volume 1, pages 598–603, October 2004.
12. A.Rajaprabu , Dr. A.Geetha, "Deep Learning Approach For Intelligent Human-Computer Interaction Using Eye Gestures", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 09-Special Issue, 2018.