IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ENHANCED COMPUTER INTERACTION USING HAND GESTURE AND SPEECH RECOGNITION

[1]Mr. R. Aravindkrishna, [2]Mr. M. Divakar, [3]Mr. S. Srisabarish, [4]Mr. S. Krishnagaanth, [5]Dr. Kalaimani Shanmugam

[1]UG Student, [2]UG Student, [3]UG Student, [4]UG Student, [5]Professor

[1]Department of Computer Science and Engineering,
[1]Arasu Engineering College, Kumbakonam, Tamilnadu

Abstract: Nowadays Hand Gesture Recognition play an important role as a technique of a development of Enhanced Human Computer Interaction. Upgraded Human Computer Interaction empowers the non-contact framework to cooperate with the PC if there should arise an occurrence of unavoidance circumstance in security and treatment. In this project, Intelligent Virtual Assistant dependent on Hand Gesture and Speech Recognition is proposed to upgrade the client cooperation. The proposed system is divided into four parts: a) hand gesture recognition; b) speech recognition; c) interface object detection; d) command execution. A deep learning method, Convolutional Neural Network(CNN) is used to extract the features of a gesture. To decide these highlights, shading division is utilized to identify the hand; shading pixels can be obtained by removing a specific HSV (tint, immersion, esteem) and applying limit covering to the input image. CNN is also used to extract feature from raw speech signal for efficient speech recognition process. Finally, a Support Vector Machine(SVM) is used to give a more accurate classification of the hand gestures, speech data and GUI Elements. So as to assess the precision of the proposed framework, the accompanying datasets are utilized: Microsoft Kinect, Leap Motion Dataset and Gesture Creative Senz3D for Gesture Detection; Google Speech Commands Dataset for Speech Recognition; pix2code dataset and Sketch2code dataset for GUI component location. The experimental results show that the proposed system can recognize gesture and speech functions with 97% accuracy.

***Index Terms* - Human-Computer Interaction(HCI), Convolutional Neural Network(CNN), Support Vector Machine(SVM), Gesture Recognition, Speech Recognition.**

I. INTRODUCTION

A non-touch system is an advanced computer interface technology used to enhance human computer interaction (HCI). The interface assists with associating a PC, machine or a robot if there should arise an occurrence of an unavoidable circumstance, security and treatment or modern life [1]. For the advancement of sensors and cameras, the non-touch devices and interfaces will continue to integrate into daily life. Lately, HCI has improved the presentation of utilization including motion acknowledgment Specifically, hand motion acknowledgment assumes a significant job as a method for the improvement of HCI, since it gives a characteristic and direct route for clients to communicate their sentiments, collaborate with computer generated reality (VR) or enlarged reality (AR) gadgets, human machine association, and to mess around. Hand gestures are the most common means for nonverbal communication.

Generally speaking, gestures are divided into two types: static gestures and dynamic gestures. The former mainly focuses on the finger's flex angles and poses while the latter pays more attention to the hand motion trajectory (HMT). In previous studies, sensors for the above two types of gesture recognition mainly referred to two categories: image-based sensors and non-image based sensors. Regarding sensory feature extraction mechanisms, deep learning approaches have recently attracted increasing attention among the machine-learning community. For example, deep neural networks (DNNs) have successfully been applied to unsupervised feature learning for single modalities such as text, images, and audio. The same approach has also been applied to the learning of fused representations over multiple modalities, resulting in significant improvements in speech recognition performance.

The remain section of the paper is organized as follows: the section II discusses about the previous work in this domain; the section III discusses the proposed work for intelligent speech and gesture recognition system; the section IV

II. RELATED WORK

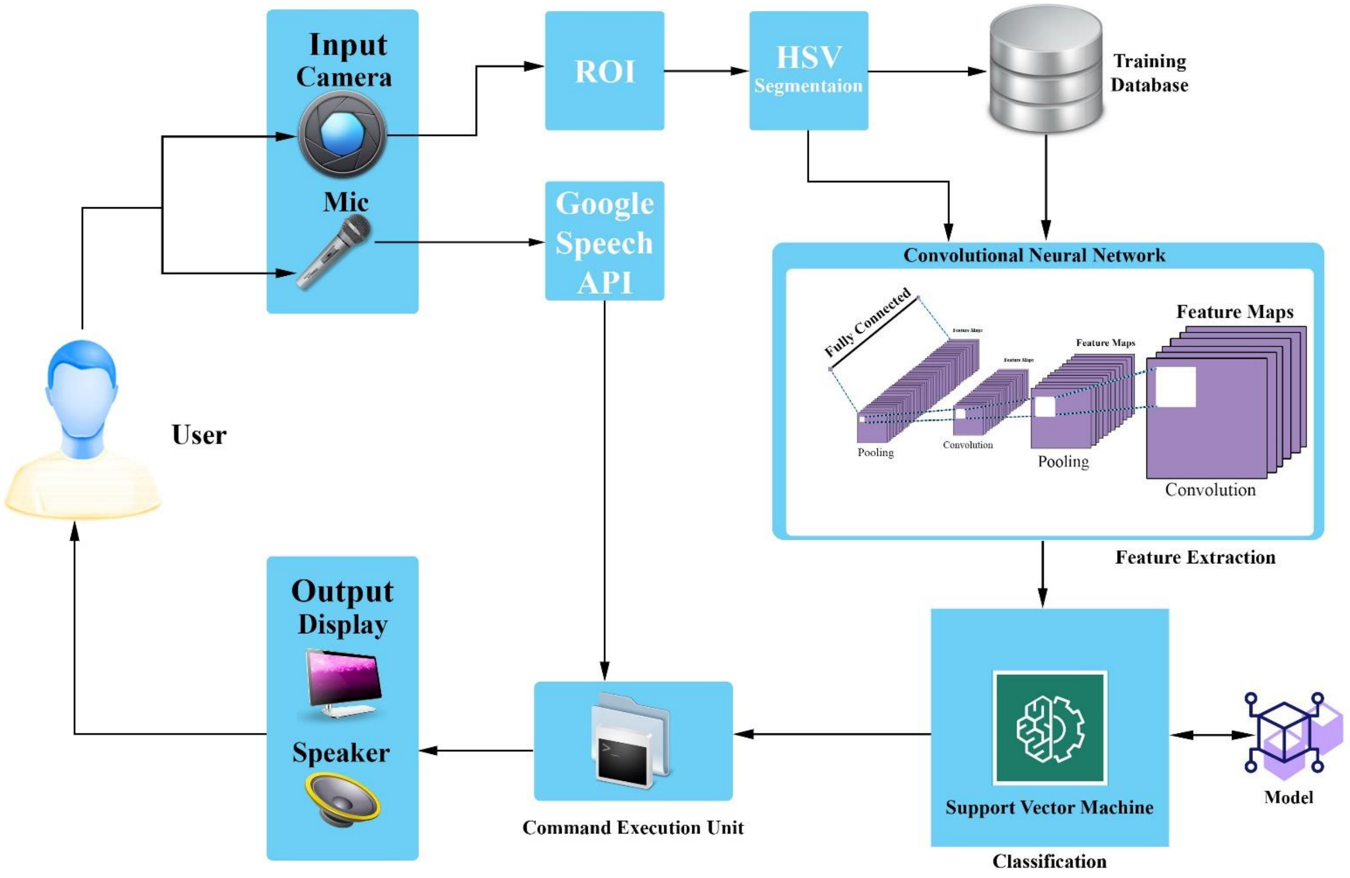
The non-touch system is a modern approach of computer-interface technology capable of revolutionizing human-computer interaction. The interface allows the user to input data and interact with a human, machine or robot in an uncontrolled environment, treatment or industrial life [1]. Information communication between two peers can be done using various mediums. These mediums can be either linguistic or gestures. The method of representing the hand gesture in binary pattern contributes a lot for increasing the performance of classification process. The binary Support Vector Machine (SVM) is considered as a recognition tool [2]. Graphical user interface (GUI) is very important to interact with software users. In many studies, therefore, they are trying to convert GUI elements (or widgets) to code or to describe formally its structure by help of domain knowledge or machine learning based algorithms. In the paper [3], object detection is adapted based on deep neural networks that finds GUI elements by integration of localization and classification. After the successfully detection of GUI components, the objects are described as the hierarchical structure and transform those to appropriate codes by synthetic or machine learning algorithms. Emerging depth sensors and new interaction paradigms [4] enable to create more immersive and intuitive

Natural User Interfaces by recognizing body gestures. One of the vision-based devices that has received plenty of attention is the Leap Motion Controller (LMC). Therefore, Hybrid Hand Gesture Recognition database, which consists of a large set of gestures generated with the LMC, including both type of gestures with different temporal sizes are introduced. Experimental results showed the robustness of the approach in terms of recognizing both type of gestures.

Based on the previous work carried out by different authors, the proposed system tries to overcome the difficulties and issues: Enhances the Interaction on various Application, Improves recognition of Patten of Gestures and Recognition on Graphical User interface.

III. Proposed System

A. System Architecture

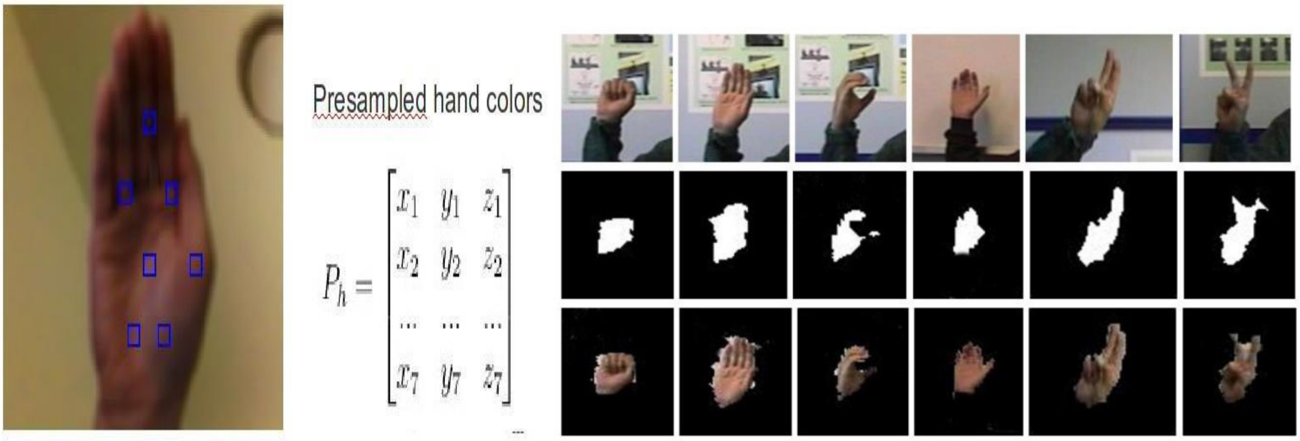


B. Preprocessing

In this module, the gesture images, GUI components images and Speech signal are pre-processed. Spectral subtraction technique is used which filter the noisy signal from the original signal by using the power of magnitude spectrum of a signal. For restoration of time-domain signals, an estimate of the instantaneous magnitude spectrum is combined with the phase of the noisy signal, and then transformed via an inverse discrete Fourier transform to the time domain. In terms of computational complexity spectral subtraction is relatively inexpensive. The processing distortion becomes more noticeable as the signal to noise ratio decreases. Hence the pre-processed Speech signal is given for Converting the Speech to text using HMM. Region of interest(ROI) he boundaries of an Input may be defined on an image or in a volume, for the purpose of measuring its size. The Hand Gesture border may be defined on an image, perhaps during different phases of the Hand Gesture. a ROI can be taken literally as a polygonal selection from a 2D map. In computer vision and optical character recognition, the ROI defines the borders of an object under consideration. In many applications, symbolic (textual) labels are added to a ROI, to describe its content in a compact manner. Hence the pre-processed image is given for training the images using CNN.

C. Convolution Layer

Convolution has a set of learnable filters which are a matrix (W x H x D). The input image is considered as a matrix and the filter is imagined sliding through the input image matrix in order to get the convoluted image which is the filtered image of the actual input image. If a filter is applied on the input image, the result would be an output matrix smaller than the original image. Padding plays an important role if we need to get the same size outputted as the input size. The purpose of convolution in image data is to extract features from the input image. Convolution will produce linear transformations of input data according to spatial information on the data. The weight on that layer specifies which convolution kernel is used, so that convolution kernels can be trained based on input on CNN.


Presampled hand colors
$P_h = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \dots & \dots & \dots \\ x_7 & y_7 & z_7 \end{bmatrix}$

Initially, the Hand Gesture of the user gesticulating (if present) is detected and removed from the second frame of the video. After the skin filtering is performed to obtain the segmented images and original images in grayscale. Using this, the palm features and the finger positions are preserved. we convert the RGB color space and use HSV color model analysis by thresholding the saturation (S) and value (V) elements. In HSV representation, the hue, saturation, and values determine the colour of the hand, the intensity of the color and the lightness of the image, respectively.

Skin coloured objects in the frame. On the other hand, three-frame differencing is performed with the first three frames. It is computed for both coloured and grayscale frames. The results of the skin filtering and three-frame differencing are combined to obtain the desired hand from the background.

1) Pooling Layer or Subsampling

In this layer, the dimensionality of the feature maps is reduced. The hand gesture image matrix obtained from the convolution is multiplied with the kernel and the pixel size is diminished here. The second layer of speech recognition is fed with some recognizable features. Here two dimensional pixels are considered. The next big part is called as pooling that is how a signal stack can be compressed. This is done by considering a small window pixel which might be a 2 by 2 window pixel or 3 by 3. On considering a 2 by 2 window pixel and pass it in strides across the filtered signals, from each window the maximum value is considered. If a pixel value is negative, then the negative values are replaced with zeros. This is done to all the filtered signals. This becomes another type of layer which is known as a rectified linear unit, a stack of signals which becomes a stack of signals with no negative values. Now the three layers are stacked up so that one output is detected. The dimensionality reduced matrix is obtained.

D. Speech Recognition

The CNN-based system is used to perform the feature learning and acoustic modelling steps, by computing the posterior probabilities of context-dependent phonemes from raw speech. The speech which is analog is converted to digital in the pre-processing phase itself. The decoder is an HMM. The scaled likelihoods are estimated by dividing the posterior probability by the prior probability of each class, estimated by counting on the training set. The hyper parameters such as, language scaling factor and the word insertion penalty are determined on the validation set. Raw features are simply composed of a window of the temporal speech signal (hence, din = 1 for the first convolutional layer). The window is normalized such that it has zero mean and unit variance. Convolutional Neural Network can do a lot of good things if they are fed with a bunch of signals for instance to learn some basic signals such as frequency changes, amplitude changes. Since, they are multi neural networks, the first layer is fed with this information. The second layer is fed with some recognizable features. Here two dimensional pixels are considered. The next big part is called as pooling that is how a signal stack can be compressed. This is done by considering a small window pixel which might be a 2 by 2 window pixel or 3 by 3. On considering a 2 by 2 window pixel and pass it in strides across the filtered signals, from each window the maximum value is considered. If a pixel value is negative, then the negative values are replaced with zeros. This is done to all the filtered signals. This becomes another type of layer which is known as a rectified linear unit, a stack of signals which becomes a stack of signals with no negative values. Now the three layers are stacked up so that one output is detected.

IV. Classification and Prediction

The hand gestures, speech and GUI elements are classified using Support Vector Machine (SVM) which is a most commonly used classifier. The gestures presented are reasonably distinct, the images are clear and without background. Also, there is a reasonable quantity of images, which makes our model more robust. The drawback is that for different problems, we would probably need more data to stir the parameters of our model into a better direction. Moreover, a deep learning model is very hard to interpret, given its abstractions. SVM classifies the data by creating the hyperplane or set of hyperplanes in a high dimensional space.

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \sigma > 0$$

V. Command Execution

This module will take the user's hand gesture and voice commands as an input. It then finds the necessary action for the given command by comparing the voice commands, GUI list and hand gesture dataset and executes the respective actions. Social command is the request response system, which is the "W" type of question asked to the system. An interactive system with the user is made by the request response system. Social command is the vast type of command whereas updating of data is a continuous process, as the requirements of the user may vary. The questions framed for the user as default are obtained from the user given requirements to the trainer. These actions are performed separately in the modules based upon the user convenience. It is done in CMU where all the input commands are executed.

VI. Dataset Descriptions

Microsoft Kinect, Leap Motion Dataset and Gesture Creative Senz3D for Gesture Detection. Google Speech Commands Dataset for Speech Recognition;

The Microsoft Research Cambridge-12 Kinect gesture data set consists of sequences of human movements, represented as body-part locations, and the associated gesture to be recognized by the system. LMDHG contains unsegmented sequences of hand gestures performed with either one hand or both hands. This dataset can therefore be used for recognition of both pre-segmented and unsegmented dynamic hand gestures using skeleton data. Each frame contains the 3D coordinates of 46 joints (23 joints for each hand). If one of the hands is not tracked, then the position of its joints are set to zero. The speech recognition dataset has 65,000 clips of one-second-long duration. Each clip contains one of the 30 different words spoken by thousands of different subjects.

VII. Conclusion and Future Enhancement

This paper proposed a hand gesture recognition model where gesture images were captured from the region of interest (ROI) of the virtual keyboard. The gesture image was segmented to detect the hand area and the proposed CNN method was used to extract the gesture features. Finally, the gestures were recognized using SVM classification techniques. The Proposed System of speech recognition started with a brief introduction of the technology and its applications in different sectors. The project part of the Report was based on software development for speech recognition. The experimental results show that our system can achieve a high recognition accuracy of 98.09% and 96.23% respectively.

References

[1] Md. Abdur Rahim & Jungpil Shin & Md. Rashedul Islam “Hand gesture recognition-based non-touch character writing system on a virtual keyboard” 2020

[2] Ashis Pradhan, Mohan Chandra Pradhan “A Hand Gesture Recognition using Feature Extraction” 2012

[3] Diego G. AlonsoEmail authorAlfredo TeyseyreLuis BerdunSilvia Schiaffino “A Deep Learning Approach for Hybrid Hand Gesture Recognition” 2019

[4] Young-Sun Yun &Junyoung Heo & Amarjit Roy “Detection of GUI Elements on Sketch Images Using Object Detector Based on Deep Neural Networks” 2019