Role of SURF and MFCC Algorithms in the Design of Human Interacting Systems

[1]Sk Nazma, [2]Lakshmi Naryana Thalluri , IE Member.
[1]M.Tech. Student, Dept. of ECE, Andhra Loyola Institute of Engineering and Technology, Vijayawada, A.P, India.
[2]Assistant Professor, Dept. of ECE, Andhra Loyola Institute of Engineering and Technology, Vijayawada, A.P, India.

Abstract— **In this paper, a new method is proposed to design Human Recognition and Interaction (HRI) System based hardware control. The HRI system first recognizes the human face, takes the speech instruction from human, and controls the interfaced hardware. Required knowledge to design this type of digital system is signal recognition. Any signal recognition system mainly deals with input signal feature detection, description, train the database, and finally matching the test signal features for classification. The signal recognition accuracy mainly depends on the strength of feature vectors. The design Process basically divided in to three phases, the first phase involves Human face recognition for that, a modified Speeded up robust features (SURF) algorithm is used to detect, describe the Human face features from input image, and Fast Library for Approximate Nearest Neighbors (FLANN) algorithm is adopted for test face classification. The second phase involves speech recognition for that, a Mel Frequency Cepstral Coefficients (MFCC) algorithm is used for feature detection, description, and Dynamic Time Warping (DTW) algorithm is used for test speech signal classification, by combining these two phases a HRI system is designed, and in the third phase the hardware components like Motor, Light is interface with HRI System through a ARDUINO UNO board. Main problem in any recognition system design is large database size, one way to reduce the database size is identifying the Region of Interest (RoI) in HRI system input signal, and this step will reduce the database size and also signal processing time. In this work, the HRI system is trained with five human faces and six isolated speech words.**

Keywords— SURF Algorithm, FLANN, MFCC Algorithm, DTW, Region of Interest.

I. INTRODUCTION

The decades of research are advanced the signal recognition area, and still there is a scope for research in signal recognition, because so many advanced applications are requiring signal recognition. Well known applications are Robotic Design, Security, and Authentication etc. As shown in Figure 1, a basic signal recognition system involves

i) Input Signal Acquisition
ii) Preprocessing
iii) Feature Detection and Description
iv) Train Database
v) Test Signal Feature Classification

In this paper, HRI system design requires two dimensional (Human Face in Image) signals, and one dimensional (Speech) signals recognition. The existed algorithms for image object feature detection, description are PCA, BRISK, BRIEF, SURF, SIFT. For object feature matching, HMM, KNN algorithm are useful. The existed algorithms for isolated speech recognition are LPC, MFCC, LPCC, and BFCC. For isolated speech feature matching, Euclidian distance, Vector Quantization, and Dynamic Time Warping (DTW) algorithms are useful.

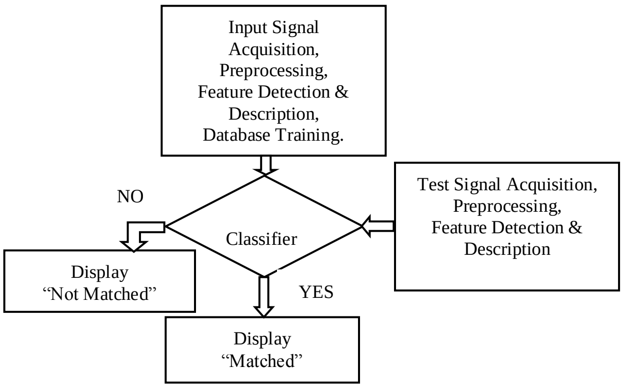

Input Signal Acquisition, Preprocessing, Feature Detection & Description, Database Training.
NO
Classifier
Test Signal Acquisition, Preprocessing, Feature Detection & Description
Display "Not Matched"
YES
Display "Matched"

Figure 1: Basic Signal Recognition System Block Diagram

Few feature detection and description algorithms Comparison is shown in table 1.

Table 1: Comparison of Different Image Feature Detectors and Descriptors [6]

Descriptor	SIFT	SURF	BRISK	BRIEF	ORG
Detector	DOG	DOH	AGAST	AGAST (or) FAST	FAST
Feature Matching Method	Flann Based	Flann Based	Brute-force Based	Brute-force Based	Brute-force Based
Rotate Invariant	Yes	Yes	Yes	No	Yes
Scale Invariant	Yes	Yes	Yes	No	No
Illumination Invariant	Yes	No (Poor)	Poor	Poor	Poor

Table 2: Comparison of Different Isolated Speech Recognition Algoritms

Algorithm	LPCC	MFCC	BFCC
Steps	1)Pre-Emphasis, 2)HammingWindowing, 3)Linear Predictive Analysis, 4)Cepstral Analysis.	1)Pre-Emphasis, 2)Frame Blocking, 3)Windowing,4)FFT, 5)Mel Scale Filter Bank, 6)Logarithmic Expression, 7)DCT	1)Pre-Emphasis, 2)Frame Blocking 3)Windowing, 4)FFT 5)Bark Scale Filter Bank, 6)Equal Loudness, 7)Intensity Loudness,Compression 8)Logarithmic Expression, 9)DCT
Frequncy Scale		$Mel_f = 2595* \log_{10}(1+f/700)$	$F_{bark} = 6* \ln(f/600+((f/600)^2+1)^{0.5})$

II. DESIGN METHODOLOGY

In this paper, a Human Recognition and Interaction (HRI) system based hardware control is designed. In the first phase the HRI system as shown in figure 2. is designed and in second phase hardware elements like motor, light are interfaced to HRI system using an interface module. Signal recognition is the required to design HRI system, signal recognition means it involves object recognition and speech recognition. In this work, the object recognition is done using modified SURF algorithm, HMM algorithm and speech recognition is done using MFCC algorithm, Dynamic Time Warping (DTW).

A. Problem Identification & Solutions:

In any recognition system design, main hurdles are achieving the accuracy is maintaining the large Database, if the size of the database is increasing required memory will increase. In this work the size of the database is reduced by introducing Region of Interest (RoI) extraction blocks in front of HRI system separately for image and speech. And one more solution is make the HRI system as robust as possible for that best key-point or feature detection and description is required with high speed. Compared to SIFT, SURF algorithm is much faster because it uses integral image for processing. But SURF features are not illumination invariant. To overcome this problem, here object (Face) feature detection and description is done by suing modified SURF algorithm, the modification is the SURF features are normalized to make the SURF features as illumination invariant.

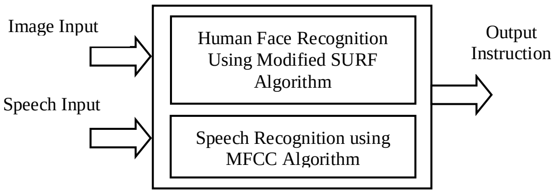


Image Input
Speech Input
Human Face Recognition Using Modified SURF Algorithm
Speech Recognition using MFCC Algorithm
Output Instruction

Figure 2: HRI System Block Diagram

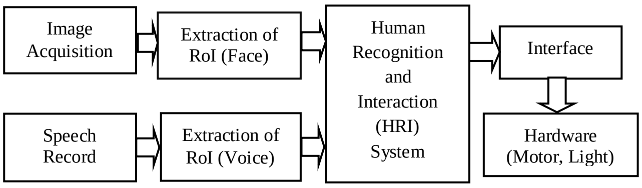


Figure 3: A HRI System Based Hardware Control Block Diagram

B. Human Face Recognition:

In object recognition *human face* detection is one area, there are so many algorithms like SIFT, SURF, BRISK, and BRIEF are existed for human face recognition. In these BRISK and BRIEF are binary algorithms, which are suitable to implement on FPGA's. Here a modified SURF algorithm is used for face recognition other than SIFT algorithm because SURF is the speeded version of SIFT algorithm.

1) Region of Interest in Image

In any recognition system, the major problem is large database size, to reduce the database size in this paper a Viola-Jones algorithm based Region of Interest (RoI) block introduced.

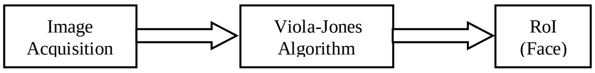


Figure 4: Human Face Detection and Extraction

2) SURF Algorithm:

For face recognition a feature detector, descriptor, and a feature matcher is required. In this a modified Speeded up robust feature (SURF) is used as a feature detector and descriptor, and a FLANN based matcher is used to match the key points. Here the modification to SURF algorithm is mainly to improve the illumination invariance of detected features. SURF algorithm is an advanced version of Scale Invariant Feature Transform (SIFT) Algorithm and it is three times faster than SIFT. SIFT uses Difference of Gaussian (DOG) based feature detection technique, it will take much time. To reduce the required time SURF algorithm uses Determinant of Hessian (DOH), it is a blob detector [10]. SURF uses integral image for description for high speed. The main advantage in SURF feature extraction is, those are scale invariant, rotate invariant, and in this paper the SURF features illumination invariance is improve by normalizing the detected Feature vector to unity. The SURF algorithm feature finding work is divided in to two steps; first step is feature detection using determinant of hessian and in the second step the detected features are described using Haar wavelet method.

1) Determinant of Hessian(DOH):

SURF use Hessian matrix as a blob detector. A point in an image I is Ip = [*x* , *y*], and the hessian matrix in I at scale σ is

$$\boldsymbol{H(x,y,\sigma)}=\begin{bmatrix} \boldsymbol{Lxx(x,y,\sigma)} & \boldsymbol{Lxy(x,y,\sigma)} \\ \boldsymbol{Lyx(x,y,\sigma)} & \boldsymbol{Lyy(x,y,\sigma)} \end{bmatrix} \qquad \text{eq. (1)}$$

Where **L*xx*(*x,y,σ*), L*xy*(*x,y,σ*), L*xy*(*x,y,σ*)** are the second order derivative to the convolution of Gaussian with image I in point Ip. The approximated hessian matrix is determined by using box filters of size 9x9, with σ =1.2 i.e.

$$\boldsymbol{H_{approx}} = \begin{bmatrix} \boldsymbol{Lxx}^{\wedge} & \boldsymbol{Lxy}^{\wedge} \\ \boldsymbol{Lyx}^{\wedge} & \boldsymbol{Lyy}^{\wedge} \end{bmatrix} \qquad \text{eq. (2)}$$

Location and scale of interesting points, SURF uses determinant of Hessian (DOH)

$$\mathbf{det}\ (\boldsymbol{H_{approx}}) =(\mathbf{L}\boldsymbol{xx}^{\wedge}\ \mathbf{L}\boldsymbol{yy}^{\wedge} - (\mathbf{0.9}\ \mathbf{L}\boldsymbol{xy}^{\wedge})^{2}) \qquad \text{eq. (3)}$$

2) Descriptor using integral image:

After localizing the interesting points in scale and space, a descriptor describes the distribution of Haar-wavelet responses within the interest point neighborhood. To make the features as rotate invariant SURF uses we first calculate the Haar-wavelet responses in *x* and *y* direction, around the interesting point in a circular shape of radius 6s, where 's' is the scale at which the interest point was detected.

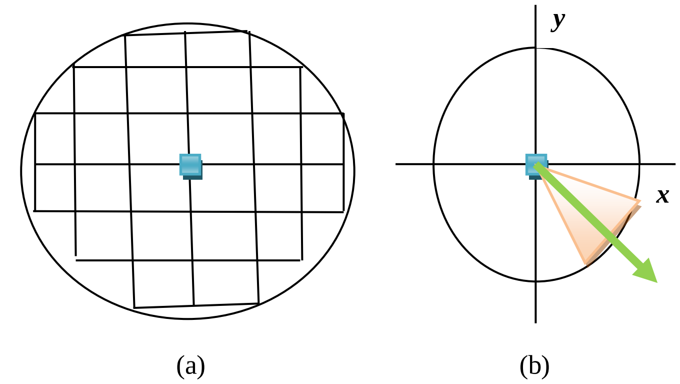


Figure 5: (a) Circular Neighbors of Interesting Point, (b) Haar Wavelets

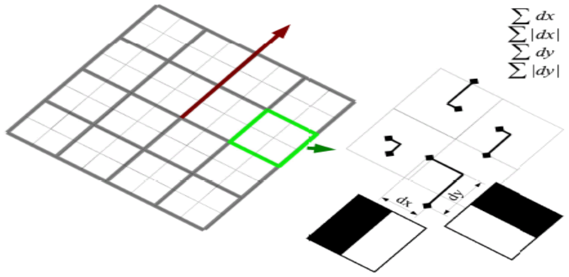


Figure 6: Key Point Description

The main modification to SURF algorithm is to improve the illuminace invariance the described feature vector is normalized to unit length here the feature description values are stored as a row vector in a feature matric so here each row in the feature matrix is normalized to unit length. Row normalization of a matrix A=[a b c] is

$$Row\ Normalization(A) = \left[\frac{a}{\sqrt{a^2+b^2+c^2}}\ \frac{b}{\sqrt{a^2+b^2+c^2}}\ \frac{c}{\sqrt{a^2+b^2+c^2}}\right] \quad \text{eq. (4)}$$

3) Feature Matchers

After describing the features, a fast and efficient feature matcher is required to match the test face features with trained features. In this paper the feature matching is done by using Fast Library for Approximate Nearest Neighbors (FLANN) based method which is suitable to match SURF features. In this paper the FLANN algorithm finds composite index i.e. it includes kdtree and kmean. 'Kdtree' creates one or more randomized kd-trees and 'kmean' creates a hierarchical kmeans clustering tree.

FLANN algorithm may implement using K-nearest neighbors.

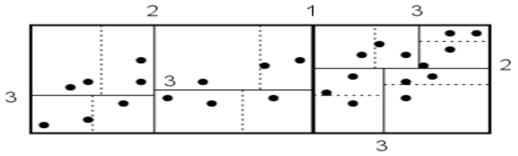


Figure 7: kd-Tree example

C. Isolated Speech Recognition:

Speech signal is a one dimensional signal, the features can extract in time domain or in frequency domain. Energy, Zero crossings, and Pitch are the time domain features. In this paper, for speech recognition frequency domain features like, cepstrum coefficients from spectrogram are considered.

1) Region of Interest in Speech:

In recorded speech always there is voice portion and non-voice portion. Processing the entire signal without separating the voice portion will require more processing time and large memory. To overcome this problem in this paper, a Magnitude window based Region of Interest (RoI) block is used to separate voice and non-voice portions due to this the processing time is reduced and required memory space is also reduced.

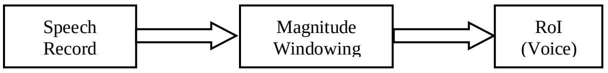


Figure 8: Voice and Non-Voice Separation

2) MFCC Algorithm:

The Mel Frequency Cepstral Coefficients (MFCC) algorithm [8] involves three steps i.e. dividing the speech signal into small windows, extraction of Spectrogram, finding the Mel-Spectrum, and finally extraction of cepstral coefficients.

At first the spectrum vectors are extracted for the input speech signal by dividing the input speech signal into short windows, preferred duration is 20ms, apply Fast Fourier Transform (FFT) to each and every window, and finally dark and white colors are mapped based on amplitude i.e. high amplitude means dark, low amplitude means white. These mapped vectors are known as spectrogram.

Mel filter banks with variable length and variable distance are applied to the Mel-Spectrum i.e. X(k) and apply logarithm for X(k). Here more number of filters is applied at low frequency region and lesser number of filters is applied at high frequency region. If the number of filters is 41 the algorithm will be effective. Mel frequency will decide the position of the filters i.e.

$$X(k)=H(k)E(k) \quad \text{eq. (5)}$$

$$Log(X(k)) = log(Mel\text{-}Spectrum)=log(H(k)) + log(E(k)) \quad \text{eq. (6)}$$

Apply cepstral analysis on log(Mel-Spectrum), to separate spectrum envelop [H(k)] and spectrum details [E(k)] i.e. spectrum envelop is a low frequency and spectrum details is high frequency. In cepstral analysis low frequency component is separated by applying inverse FFT to log(X(K)) and passed it through high pass filter and low pass filter under Pseudo frequency axis. x(k) is referred as cepstrum.

$$IFFT(\log X(k)) = x(k) = h(k) + e(k) \qquad \text{eq. (7)}$$

here h(k) is the spectral envelop, which is used as features for speech recognition. Why mel scale frequency means, mel scale frequency is approximate to human hearing perceptional frequency. An expression to convert in to Mel-scale is

$$Mel_f = 2595 * \ln(1+f/700) \qquad \text{eq. (8)}$$

3) Dynamic Time Warping (DTW):

Dynamic time warping is the technique to find the similarity between two sequences with different scales [11]. It will take the two set of sequences in to consideration and compare each and every sample in one sequence with the samples in the other sequence.It is an optimal match optimal match between two sequences with certain restrictions. For example, similarities in walking patterns can be measured by using DTW.

As an example, here two number sequences X & Y are matches by using DTW, i.e. X = (1,1,4,2,5) and Y= (1,2,3,2,2,4,6)

6	5	5	2	4	1
4	3	3	0	2	1
2	1	1	2	0	3
2	1	1	2	0	3
3	2	2	1	1	2
2	1	1	2	0	3
1	0	0	3	1	4
	1	1	4	2	5

Figure 9: Tabular Representation to Find Distance Between X & Y

In the above figure we can find the shortest distance between two sequences of different length i.e. shortest distance = 4, which cannot be possible in Euclidean method.

D. Hardware Interface

A programmable Logic Device (PLD) is serially interfaced to HRI system with a baud rate of 9600. Here the PLD (Arduino Board) will get the instructions from the HRI system, based on that hardware connected to the PLD will operate.

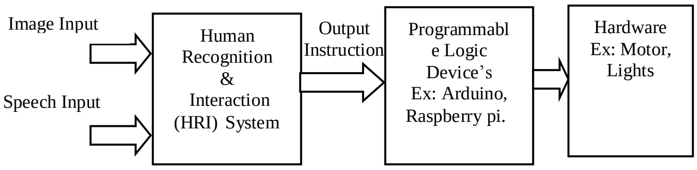


Image Input
Speech Input
Human Recognition & Interaction (HRI) System
Output Instruction
Programmable Logic Device's Ex: Arduino, Raspberry pi.
Hardware Ex: Motor, Lights

Figure 10: HRI System Interface with Hardware

III. RESULAT ANALYSIS

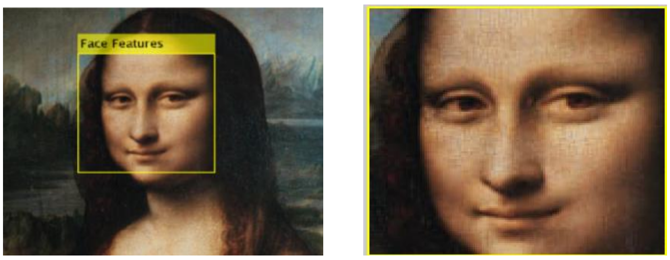
The Human Recognition and Interaction system design involves face, isolated speech detection and recognition. For this two databases need to train one is for face detection, recognition and another one is for Speech detection and recognition. The complete project is done on MATLAB 2015a.

A. Face Detection & Recognition:

In the first phase face detection and recognition system is designed. For face detection face features are extracted and trained using Viola-Jones object recognition algorithm [3]. After that for face recognition another database is created with SURF features.

1) Face Detection:

In any signal recognition system design require trained database, if it is speech recognition, database need to train with different speech signal. Similarly if it is object(face) recognition system require a well trained object database. and here main problem is size of the database. But accuracy and database size both are proportional, i.e. for accuracy the database needs to train with large number of signal. In this paper, the database is trained only with detected faces from input image. Viola-jones object recognition algorithm [3] is used to detect the face. It will detect the face in the input image as shown in figure 10(a), the detected face is cropped as shown in figure 10(b) and fatherly used for database training Due to this the database size is reduced significantly.



(a) (b)

Figure 11: Extraction of ROI (a) Detected Face Features (b) Cropped Face.

2) Face recognition using SURF:

For face recognition, the face database is trained with speeded up robust Features (SURF), because these features are rotate invariant, scale invariant, and speed is high compared with scale invariant feature transform(SIFT) algorithm. Here the extracted features are scale invariant, rotate invariant, and because of normalization to unit length these features will be illumination invariant. Because of this the modified shift features are most preferable for real time applications.

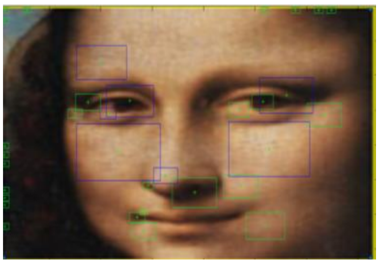


Figure 12: SURF Features

3) Feature Matching:

The main step in signal recognition system is matching of Features. There are different approaches are exist; in this a FLANN Based matched is designed as a Feature matcher. Which was used to find the index with **kdtree** and **kmean**. The feature matching is verified with same faces indicates "Matched" in the first case shown in Figure 12, and in the second case feature matching is done with different faces indicates "Not Matched " as shown in Figure 13.

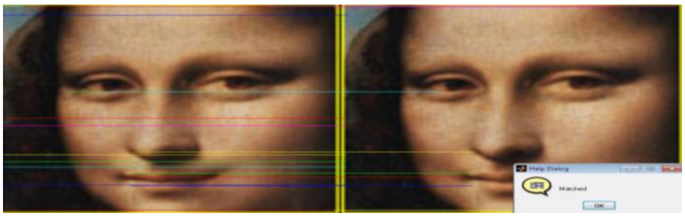


Figure 13: Feature matching of same faces indicates "Matched".

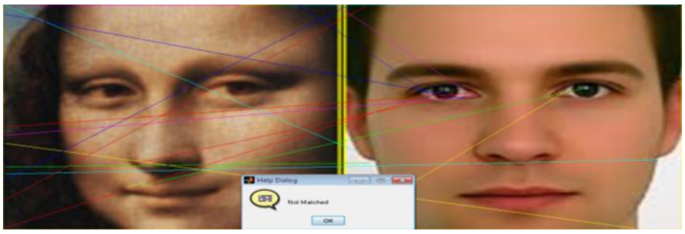


Figure 14: With Different faces indicates "Not Matched".

B. Speech Detection & Recognition:

The input speech is recorded for the duration of 2 seconds with a sampling rate of 8000 Hz and each sample is defined with 16 bits. Initially the input speech is combination of voice and non-voice.

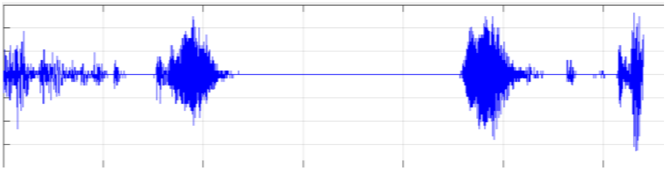
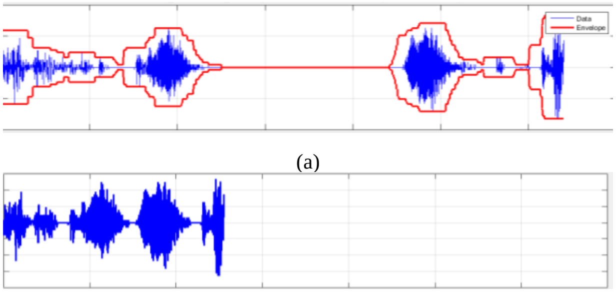


Figure 15: Speech Combination of Voice and Non-Voice Portions

The unwanted non-voice potion is removed to get the region of interest by using envelop detector, as shown in figure 16.

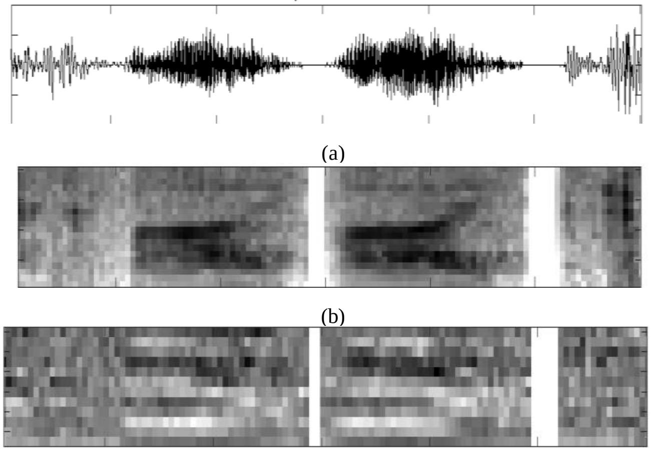


(a)

(b)

Figure 16: Separation of Voiced & Non-Voiced Portions (a)Voice & Non-Voice Signal, (b)Only Voice Signal.

After getting the pure voice signal the MFCC features are extracted as shown in figure 17.



(a)

(b)

(c)

Figure 17: MFCC Extraction (a) Speech Wave, (b) Log(mel) Filterbank Energies, (c) Mel Frequency Capstrum.

IV. CONCLUSION

This paper mainly deals with the design and simulation of Human Recognition and Interaction system based hardware control. SURF algorithm is used for face features extraction, and MFCC algorithm is used to extract the speech features. The major advancement in this paper is extraction if region of interest from input signal i.e. image and speech. Because of this the size of the database is reduced significantly. And additionally the SURF Features are normalized to improve the luminance invariance. After completion of the Human recognition and Interaction system design, the system is interfaced with hardware like motor and light by establishing serial communication with 9600 baud rate.

V. REFERENCE

[1] Speech synthesis techniques. A survey Tabet, Y. ; Univ. of M"hamed Bouguerra Boumerdes, Boumerdes, Algeria ; Boughazi, M., Published in: Systems, Signal Processing and their Applications (WOSSPA), 2011 7th International Workshop on Date of Conference: 9-11 May 2011. Page(s): 67 – 70 Print ISBN: 978-1-4577-0689-9 INSPEC Accession Number: 12085216

[2] Leutenegger, Stefan, Margarita Chli, and Roland Y. Siegwart. "BRISK: Binary robust invariant scalable keypoints." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.

[3] "Robust Real-Time Object Detection" ,Paul Viola ,Michael Jones, Second International Workshop On Statistical And Computational Theories Of Vision – Modeling, Learning, Computing, And Sampling Vancouver, Canada, July 13, 2001.

[4] Ying Hu And Guizhong Liu, Member, IEEE, "Separation Of Singing Voice Using Nonnegative Matrix Partial Co-Factorization For Singer Identification" IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 4, APRIL 2015.

[5] S. Umesh, Member, IEEE, And Rohit Sinha, Member, IEEE," A Study Of Filter Bank Smoothing In MFCC Features For Recognition Of Children's Speech", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 8, NOVEMBER 2007.

[6] Lakshmi Narayana Thalluri; P. Bosebabu; S R Sastry Kalavakolanu; G Roopa Krishna Chandra "Design of human face detection and recognition system along with speech synthesis subtitle" Published in:ICCICCT,Date of Conference:18-19 Dec. 2015, Page(s):396 - 400INSPEC, DOI:10.1109/ ICCICCT.2015.7475311, Publisher:IEEE.

[7] R. Ramos-Lara, M. López-García, E. Cantó-Navarro, and L. Puente-Rodriguez, "Real-time speaker verification system implemented on reconfigurable hardware," J. Signal Process. Syst., vol. 71, no. 2, pp. 89–103, May 2013.

[8] Jihyuck Jo, Hoyoung Yoo, and In-Cheol Park, "Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems," IEEE Transactions on Very Large Scale Integration (VLSI) Systems,March 2015.

[9] Seongyong Ahn Hyejong Hong , Hyunjin Kim , Jin-Ho Ahn , Dongmyong Baek and Sungho Kang "A hardware-efficent multi-character string matching architecture using brute-force algorithm", SoC Design Conference (ISOCC),2009.

[10] Abdul Jabbar Siddiqui ; "Real-Time Vehicle Make and Model Recognition Based on a Bag of SURF Features" IEEE Transactions on Intelligent Transportation Systems (Volume:PP , Issue: 99) Page(s): 1 – 15 ISSN : 1524-9050,DOI: 10.1109/TITS.2016.2545640 Date of Publication : 25 April 2016.

[11] H. Kim ; Korea University, Republic of Korea ; J. Sa ; Y. Chung ; D. Park more authors ,"Fault diagnosis of railway point machines using dynamic time warping", Published in: Electronics Letters (Volume:52 , Issue: 10) Page(s): 818 – 819 ISSN : 0013-5194, INSPEC Accession Number: 15953853, Date of Publication :5 12 2016 Publisher:IET.