ISSN: 2321-7782 (Online)
Volume 2, Issue 3, March 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

Eye and Speech Fusion in Human Computer Interaction

Madhuri R. Dubey[1]
Student, M.Tech CSE
G.H.Raisoni College of Engineering
Nagpur – India

S. A. Chhabria[2]
HOD, IT Department
G.H.Raisoni College of Engineering
Nagpur – India

Abstract: This paper highlights the concept of eye and speech fusion in human computer interaction as human uses wide variety of senses to express or commands for variable activity. The proposed system considers mainly three modules which are as eye recognition, speech recognition and fusion, in addition to that different fusion techniques are applied on multimodal senses like eye and speech for certain desktop application. The feature extraction and recognition techniques for eye tracking and speech detection are also explained. The main aim of this paper is to implement various fusion techniques and evaluate the optimised fusion technique with respect to speed, accuracy, and efficiency.

Keywords: Human computer interaction, Fusion, multimodal senses, Feature extraction, optimised fusion

I. INTRODUCTION

Humans interact with computers through a user interface which includes software, such as what is displayed on the computer monitor, and hardware, such as the mouse, keyboard and other peripheral devices. Human computer interaction aims to enhance the interactions between users and computers by making computers more interactive and receptive to users' needs [1]. HCI used to design systems that minimize the barrier between the subjective natures of human needs and the computer's understanding of the user's task.

Speech and eye are important modes of senses in human-human and human–computer interactions. Speech signals provide valuable information which is required for understanding human activities and interactions; also eye frames needed real time video tracking which provides interaction of machine with different movement of pupil. Human activity is basically captured using lists of audio and visual sensors like camera, microphone. Human uses a variety of modes of information like audio, visual, touch to recognize people and understand their activity, and hence the fusion of multiple sources of information is a mechanism to robustly recognize human activity and intent in the context of human computer interaction.

II. SIMILAR WORK

- There are various techniques used for the eye tracking are as follows:

a. Template matching method

This method compares user template with templates from the database using a matching algorithm. The matching metric will evaluate similarity measure between two iris templates. This metric gives a range of values by comparing templates from the same iris, and another range of values by comparing templates from different irises. [2]

The major drawback of this method is that they need a large training sample set, which must be prepared in advance into database. The training time is very long and the results strongly depend on the training set.

b. Pulling and pushing method

Pulling and pushing (PP) procedure has been developed in order to accurately localize the circular iris boundaries. The PP method directly finds the shortest path to the valid parameters. It used to remove the reflection in order to exclude the specularities involved in the input images, also it uses an Adaboost-cascade pupil detector detect the pupil and also to exclude the non pupil image parts before further processing such that redundant computations can be avoided.[3]

The main difficulties with this approach are sharp irregularity of eyelids and segmentation error. Also, the eyelashes as well as the shadows have been detected with statistically learned prediction model which is complex.

c. Phase based method

The phase based method recognize iris patterns based on phase information which is independent of imaging contrast and illumination. This method was performed using boolean XOR operator applied to 2048 bit phase vectors to encode any two iris patterns, masked (ANDed) by both of their corresponding mask bit vectors. From the resultant bit vector and mask bit vectors, the dissimilarity measure between any two irises patterns is computed using Hamming Distance (HD) as, iriscodes are different for two different samples.

The recognition in this method is the failure of a test of statistical independence involving degrees of freedom. [4]

The subset of data with large pupils showed worst performance with higher error rate. Also, visibility in the iris area is reduced and greater part of iris is occluded by eyelids which provide less information for iris code generation and due to which success rate slows down.

d. Texture analysis method

Wildes proposed iris recognition based on texture analysis. In this method, the limbus and pupil are modelled with circular contours which are extended to upper and lower eyelids with parabolic arcs. The matching is based on normalised correlation between the acquired and database images. Classification is performed using Fisher’s linear discriminant function. The iris code was produced using wavelet packets. The whole image is analyzed at different resolutions. [5]

The main drawback of this approach is that computation of threshold for filtering and segmentation is not easy.

➢ There are various techniques used for the speech recognition are as follows:

a. Pattern Recognition approach

The pattern-matching approach has become the predominant method for speech recognition as it involves two essential steps as, pattern training and pattern comparison. Pattern training is processed using consistent speech pattern representations, for reliable pattern comparison. A speech pattern can be represented in the form of a speech template or a statistical model i.e. in hidden markov model and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern comparison stage, a direct comparison is made between the unrecognized speeches with each possible pattern learned in the training stage.

But it has disadvantage that it requires high Maintainance of Database which stores Pattern requires.

b. Knowledge based approaches

Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. An expert knowledge about variations in speech is hand coded into a system. This has the advantage of explicit modelling variations in speech; but unfortunately such Vector Quantization (VQ) is often applied to automatic speech recognition (ASR). In this approach, each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The speech signal is evaluated by all codebooks and computed the lowest distance measure.

The expert knowledge is difficult to obtain and use successfully. Thus this approach was judged to be impractical and automatic learning procedure was sought instead.

c. Dynamic time warping (DTW)

Dynamic time warping is an algorithm for measuring similarity between two sequences with respect to time or speed. Basically, DTW is a method that allows a computer to find an optimal match between two given speech sequences. The sequences are "warped" nonlinearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models. Dynamic time warping compresses various sections of utterance so as to find alignment that result in best possible match between template and utterance on frameset basis. [6]

But it has limitation that, it doesn't support optimization process using dynamic programming in DTW.

d. Statistical based approaches

In this approach variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models. Also, the word models were constructed for combining phonetic and fenonic models using K-means algorithm which is used for statistical and clustering algorithm of speech Based on the attribute of data. In this algorithm clustered the vectors based on attributes into k partitions. [6] [7]

The main disadvantage of statistical models is that they must take priori modelling assumptions which are answerable to be inaccurate, handicapping the system performance.

III. System Architecture

3.1 Eye Recognition Module

The eye recognition module encompasses two basic stages which are feature extraction of eye and eye recognition. There were various techniques for feature extraction like template-based, colour based feature extraction, delauncy triangulation, Appearance-based Approaches, etc…

But geometry feature extraction is more suitable to this project. [8]

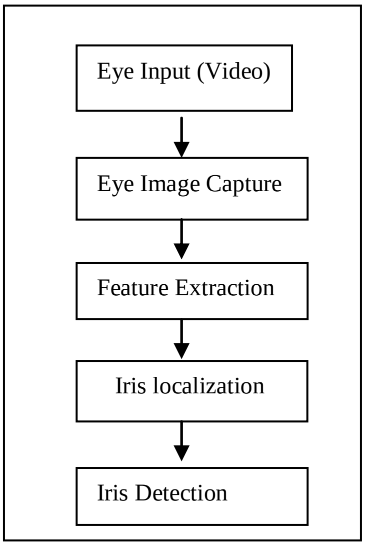


Fig 1. Eye Tracking Module

1. Input image is binarized with threshold.
2. Extract the eye based on two eye corners that are selected from corners of binary image obtained in step 1.
3. Find contours of the eye image, and then select the contour that contains iris boundary.

4. Matching between the iris boundary model and the contour obtained from step 3 to detect iris position.
5. Locate the centre of pupil by subtracting contour coordinate from actual eye coordinate which gives distance between iris boundaries to centre.
6. Repeating step 1 to step 4 to get accurate pupil features from input eye frameset.

The eye recognition is processed using Kalman filter and 2d-measurement method under motion based approach. [9]

1. The rough location of eye is localized by horizontal projection of the gradient image using Kalman Filter, 2D measurement as;
2. calculate the partial derivatives into time t and spatial directions of the image domain for each pixel within the localized eye region for two consecutive frames of a video sequence defined as;

   $Ix(x; y; t) = I(x + 1; y; t) - I(x + 1; y; t);$

   $Iy(x; y; t) = I(x; y + 1; t) - I(x- y + 1; t);$

   $It(x; y; t) = I(x; y; t + 1) - I(x-y; t + 1);$
3. calculate a linear combination of the vectors with weighting coefficients equal to absolute value of the components, and obtain two weighted vectors
4. Calculate the difference between the y-coordinates of vectors.
5. Apply ellipse detection technique to detect the precise location of eye.

3.2 Speech Recognition Module

The speech recognition module also contains speech feature extraction and speech recognition. Features of speech can be extracted using methods like, Mel-frequency cepstral coefficients (MFCC), Perceptual linear prediction (PLP), linear predictive coding (LPC), Principal component analysis (PCA), etc…

But discrete wavelet transform (DWT) is appropriate method of extracting speech features by means of transforming speech signal to waveform. [10]

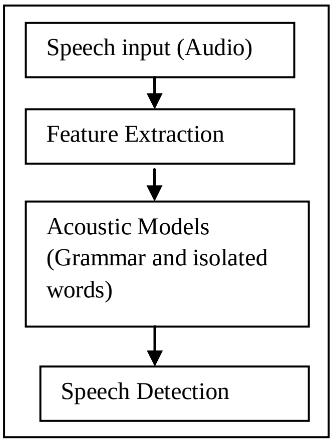


Fig 2. Speech Recognition Module

1. Speech signals are non-stationary in nature, the temporal information is also important for speech recognition; temporal information is obtained by re-scaling and shifting an analysing mother wavelet.
2. The input speech signal is analysed at different frequencies with different resolutions.
3. The DWT implementation consists of dividing the speech signal under test into approximation and detail coefficients.

4. The DWT coefficients of the input speech signal are obtained by concatenating the approximation and detail coefficients.

The speech recognition is processed using Acoustic-phonetic approach which uses language model, isolated words,

3.3 Fusion Module

Multimodal fusion is the combination of various modalities, like eye, speech, touch, hand, etc..,

The eye and speech fusion can be implemented in following manner as shown;

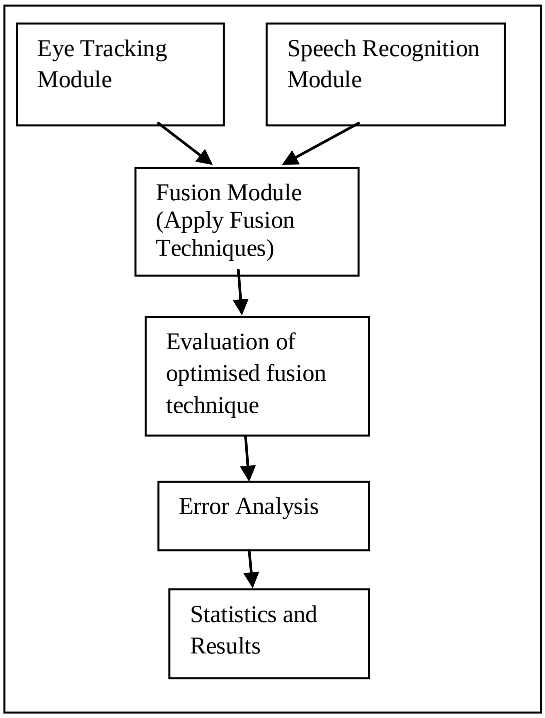

Eye Tracking Module
Speech Recognition Module
Fusion Module (Apply Fusion Techniques)
Evaluation of optimised fusion technique
Error Analysis
Statistics and Results

Fig 3. Eye – Speech fusion Module

Fusion can be taken place either before matching input sample or after matching, hence before matching fusion can be categorised as[18] [19] ;

1. Feature level fusion
2. Sensor level Fusion

- **Feature level** fusion is combination of different feature sets extracted from multiple sources of senses. Feature sets may be either homogeneous or heterogeneous.
- **Sensor level** fusion is applicable if and only if multiple sources represent samples of the single source of sense obtained either using a single sensor or different compatible sensors.

After matching, fusion can be stated as;

1. Dynamic classifier selection
2. Classifier fusion

   2.1 Score level

   2.2 Rank level

   2.3 Decision level

Score level fusion provides richest set of information. This fusion use scores from different modalities based on different scaling methods, the scores cannot be combined or used directly. It is required to perform convert the scores into common domain or scale.

Rank level fusion uses ranks output by the individual subsystems in order to derive a rank of each identity of sense. Rank level fusion provides less information as compare to score level fusion.

Decision level fusion is carried out at decision level when the decisions output by singular or multiple modes of samples are available.

IV. Scope

The use of eye and speech recognition is in its infancy, as it only works best with images and videos taken in closed room.

- It has key challenge to improve accuracy of the system to recognize all suited conditions that affect system liability.
- Gestural commands through modes of senses like eye, speech, etc.., mainly used in the gaming context. But in future, it will involve larger groups of users interacting simultaneously with reference to the commands of these modes of senses and multimodal fusion.
- This application will be needed to recreate wire frame models of body motion and vector-based dynamics (for speed and direction of movement).
- The fusion of multiple modes of senses will meet the timeliness by minimizing the re-computation of inputs and maps various function like virtual reality, augmented reality in future aspect.
- This fusion is more beneficial to social cause as it can address to disabled people.

V. Conclusion

This paper provides review of various recognition and feature extraction techniques for eye and speech; also it implemented the better suited technique of recognition for the designed application of multiple modes of senses. It specifies various fusion techniques which are applied on proposed application to find optimised and error free fusion technique with respect to the given application.

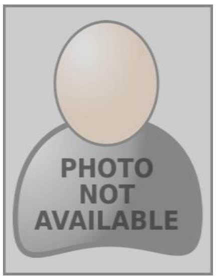
References

1. Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh and Mo Nours Arab, “Human-Computer Interaction: Overview on State of the Art”, International journal on smart sensing and intelligent systems, vol. 1, no. 1, March 2008.
2. Krystian Radlak, Bogdan Smolka, “A Novel Approach to the Eye Movement Analysis Using a High Speed Camera,” 2nd International Conference on Advances in Computational Tools for Engineering Applications 2012.
3. Shankar T. Shivappa, Bhaskar D. Rao, and Mohan Manubhai Trivedi, “Audio-Visual Fusion and Tracking with Multilevel Iterative Decoding: Framework and Experimental Evaluation,” IEEE Journal of signal processing, vol. 4, no. 5, October 2010.
4. Sruthi.T.K, “Literature review: Iris Segmentation Approaches for Iris Recognition Systems,” International Journal of Computational Engineering Research Vol, 03Issue, 5, May 2013.
5. S V Sheela, P A Vijaya “Iris Recognition Methods - Survey”, International Journal of Computer Applications Volume 3 – No.5, June 2010.
6. Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar “A Review on Speech Recognition Technique”, International Journal of Computer Applications Vol. 10–issue 3, November 2010.
7. Alejandro Canovas, Jesus Tom´as, Jaime Lloret, Miguel Garc´ıa, “Statistical Speech Translation System based on Voice Recognition Optimization using Multimodal Sources of Knowledge and Characteristics Vectors”, doi: 10.1016/j.csi. 2012.09.003, Sept 2012.
8. Nguyen Huu Cuong, Huynh Thai Hoang , “Eye-Gaze Detection with a Single WebCAM Based on Geometry Features Extraction”, IEEE 978-1-4244-7815-6/10/$26.00 ©2010.
9. S. A. Quadri and Othman Sidek,” Pixel-Level Image Fusion using Kalman Algorithm”, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 6, No. 2, April, 2013.
10. Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, “Comparative study of automatic speech recognition techniques”, IET Signal Processing Vol. 7, Iss. 1, 2013.

11. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, " Multimodal fusion for multimedia analysis: a survey", Multimedia Systems, Springer- 2010.
12. Norman Poh and Josef Kittler, "A Unified Framework for Biometric Expert Fusion Incorporating Quality Measures", IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 1, January 2012.
13. Wesley Mattheyses, Lukas Latacz and Werner Verhelst, "Multimodal Coherency Issues in Designing and Optimizing Audiovisual Speech Synthesis Techniques" International Conference on Audio-Visual Speech Processing University of East Anglia, Norwich, UK, September 2009.
14. Mattheyses, W., Latacz, L., Verhelst, W. and Sahli, H, "Multimodal Unit Selection for 2D Audiovisual Text-to-Speech Synthesis", Springer Lecture Notes in Computer Science, Volume 4261 125–136, 2008.
15. Norman Poh and Josef Kittler, "On Using Error Bounds to Optimize Cost-Sensitive Multimodal Biometric Authentication"
16. Chang-Hwan Im, Hong-Kyu Kim, Hyun-Kyo Jung, "A Novel Algorithm for Multimodal Function Optimization Based on Evolution Strategy", IEEE transactions on magnetics, vol. 40, no. 2, march 2004.
17. Mohamed Soltane and Mimen Bakhti, "Soft Decision Level Fusion Approach to a Combined Behavioral Speech-Signature Biometrics Verification", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 6, No. 1, February, 2013.
18. Satoshi Tamura, Koji Iwano and Sadaoki Furui, "Toward robust multimodal speech recognition".
19. Nirmalya Roy, Sajal K. Das and Christine Julien, "Resource-Optimized Quality-Assured Ambiguous Context Mediation Framework in Pervasive Environments", IEEE transactions on mobile computing, vol. 11, no. 2, February 2012.
20. Joyce Y. Chai Zahar Prasov, Pengyu Hong, "Performance Evaluation and Error Analysis for Multimodal Reference Resolution in a Conversation System"
21. Matthias Wimmer, Bj¨orn Schuller, Dejan Arsic, Gerhard Rigoll, Bernd Radig "Low-level fusion of audio and video feature for multi-modal emotion recognition" IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 2008.
22. Dapindar Kaur, Gaganpreet Kaur, "Level of Fusion in Multimodal Biometrics", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 2, February 2013.
23. Annis Fathima, S.Vasuhi, N.T.Naresh Babu3, V.Vaidehi, Teena Mary Treesa, "Fusion Framework for Multimodal Biometric Person Authentication System", IAENG International Journal of Computer Science, 41:1, IJCS_41_1_02, February 2014.
24. Sudhamani M J, M K Venkatesha, K R Radhika," Revisiting Feature level and Score level Fusion Techniques in Multimodal Biometrics System", 978-1-4673-1520-3/12/$31.00 IEEE 2012.
25. Puente Rodríguez, A. García Crespo, M. J. Poza Lara, B. Ruiz Mezcua, "Study of Different Fusion Techniques for Multimodal Biometric Authentication", IEEE International Conference on Wireless & Mobile Computing, Networking & Communication.
26. Divyakant T. Meva, C. K. Kumbharana, "Comparative Study of Different Fusion Techniques in Multimodal Biometric Authentication" International Journal of Computer Applications Volume 66– No.19, March 2013.

AUTHOR(S) PROFILE

Madhuri Rajnarayan Dubey is pursuing Master of technology in Computer Science Engineering discipline from G. H. Raisoni college of Engineering, Nagpur, Maharashtra.



S.A. Chhabria is currently working as professor in Information Technology department, G. H. Raisoni college of Engineering, Nagpur, Maharashtra.