# Real Time Fusion of Gestures

Sharda Chhabria
*Department of Computer Science Engineering*
*GHRCE, NAGPUR, India*

Dr. R.V.Daraskar
*Former Director,*
*Disha Education Society,(DIMAT*
*- Disha Technical Campus)*
*Raipur, India.*

Dr. V.M.Thakare
*Professor & Head,CSE,*
*Amravati University, India*

**Abstract- This paper presents a real-time fusion of eye and speech gesture recognition method. Such fusion of gesture is very useful / beneficial for different kind of Handicapped / Peoples.The gesture recognition from the video sequences is one of the most important challenges in the computer vision. It offers to the system, the ability to identify, recognize and interpret the human gestures in order to control some devices.**

**The increasing availability of sensors able to provide real-time depth measurements, such as time-of-flight cameras or the more recent Kinect, has helped researchers to find more and more efficient solutions for these issues. With the main aim to implement effective gesture-based interaction systems, this study presents an fusion approach of eye and speech that exploits two different video streams: eye and speech.. The entire procedure is designed to maintain a low computational cost and is optimized to efficiently execute HCI tasks.This paper discusses the need of gesture fusion and also the methodologies used are AND/OR fusion and Sum based fusion,quality score based fusion methods .Performance is compared in terms of accuracy and error rate.**

**Keywords – Real-time gesture recognition;Fusion,optimization,efficiency.**

## I. Introduction

As computers become more pervasive in society,facilitating natural human–computer interaction (HCI) will have a positive impact on their use. Hence, there has been growing interest in the development of new approaches and technologies for bridging the human–computer barrier. The ultimate aim is to bring HCI to aregime where interactions with computers will be as natural as interactions between humans, and to this end, incorporating gestures in HCI is an important research area.We are interested in developing a vision-based system which can interpret a user's gestures in real time to manipulate windows and objects within a graphical user interface (GUI).

## II. NEED OF GESTURE FUSION

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. The integration of multiple media, their associated features, or the intermediate decisions in order to perform an analysis task is referred to as multimodal fusion. Humans are having various mode of senses through which interaction, communication, etc.., can be processed. The senses like speech and eye are important modes to carry human-human and human–computer interactions [1]. Speech signals provide valuable information which is required for understanding human activities and interactions using voice commands; also eye frames needed real time video tracking which provides interaction of machine with different movement of eye. The combination of input from various modes of senses enables the development of human intelligent systems known as fusion of multiple modalities.

Humans may process information faster and better when it is presented in multiple. Some of the key advantages [2] of multiple modes fusion are described as below;

- They permit the flexible use of input modes, including alternation and integrated use.
- They can support shorter and simpler speech utterances than a speech-only interface, which results in fewer disfluencies and more robust speech recognition.
- They lead to enhanced error avoidance and ease of error resolution.
- They accommodate a wider range of users' response, tasks, and variable situations.
- They accommodate individual differences, such as permanent or temporary handicaps also helpful for social cause.
- They can help to prevent overuse of any individual mode during extended computer usage of hardware and other component.

## III. LITERATURE REVIEW

Sicong Zheng, (2010) [3], motivated by the potential and promise of image fusion technologies in the multi sensor

The users interaction with computers through multiple modalities such as speech, gesture, and gaze is explained in Bolt 1980; Cassell et al., 1999; Cohen et al., 1996; Chai et al., 2002; Johnston et al., 2002. There are various optimization techniques available in multimodal fusion proposed in the set of papers, some are reviewed below;

- Reference resolution Technique [3] is used to find the most proper referents to referring expressions. This technique focussed on graph matching algorithm.The main aim of this technique is to find a match between the referring graph and the referent graph that achieves the maximum compatibility between the two graphs. This method is optimised as for complex input with multiple referring expressions was considered correctly and it resolved only if the referents to all the referring expressions were correctly identified, but it has some technological limitation like disfluencies in speech utterances, and variation in the input quality or the environmental condition may hamper the real-time performance seriously.

- Optimal coupling Method [4] states that When Fusion of two audiovisual segments are involved, then audio sample and a video frame will be selected first, the fused points are referred as cutpoints. In a first approach, the known fact is that humans are highly Sensitive towards the audio track as compared to the video track. But exceptionally, the lead of the visual speech in front of the auditory speech exists. But this approach causes a minimal desynchronization between the fusion of audio track and video track when the sequences of audio video segments are already joined. In second approach, the set of probable end frames and start frames are selected in the plane of audio mode. Secondly, one frame from each set is chosen as final cutpoint, based on the minimization of the visual join cost calculated for every level of fusion of end frame-start frame.But this technique will cause extra desynchronization of fusion of audio and video track, since there will be an increased and varying difference between the video cutpoints and the audio cutpoints fused at certain level.

## IV. METHODOLOGY USED

- Fusion at the Matcher Score Level:
- At this level, the match scores output by multiple experts are combined to generate a new output (a scalar or vector) that can be subsequently used for decision-making. Fusion at this level is the most commonly discussed approach primarily due to the ease of accessing and processing match scores (compared to the raw data or the feature set extracted from the data). Fusion methods at this level are described as below:[5]

Fixed rules

- *. AND fusion*

In AND fusion, the outputs of different classifiers are threshold. An acceptance decision is reached only when all the classifiers agree.

- *OR fusion*

In OR fusion, again the outputs of different classifiers are compared to a preset threshold. A positive decision is made as soon as one of the classifiers makes an acceptance decision.

- Sum Rule-based Fusion:

  The procedure for sum rule-based fusion [6] is stated as set of normalized scores (x1, x2, .., xm) from a particular person with the index i = 1, 2,……..m. the fused score fs is evaluated using the formula

  fs = w1 x1 + w2 x2 + ... + wm xm;

  The notation wi stands for the weight which is assigned to the matcher 'i', for i = 1,...m.
  In the next step, the fused score fs will be compared to a pre-specified threshold t. We declare that a person to be a genuine user if fs > t, otherwise, sample does not match declared as impostor.

  The Sum Rule based fusion algorithm [1] is explained as shown below;

  Step1. Let {Xi,Yi} and {Xj ,Yj} be vectors obtained at two different time instances i and j. Here, X and Y represent the feature vectors derived from two different information sources.

  Step2. Let sX and sY be the normalized match scores generated by comparing Xi with Xj and Yi with Yj , respectively,  match score obtained using the simple sum rule.

  n_iris_MM = 1/ (max(IrisGI)-min(IrisGI)) * (IrisGI-min(IrisGI));
  n_Speech_MM=1/(max(SpeechGI)-min(SpeechGI))*(SpeechGI-min(SpeechGI));

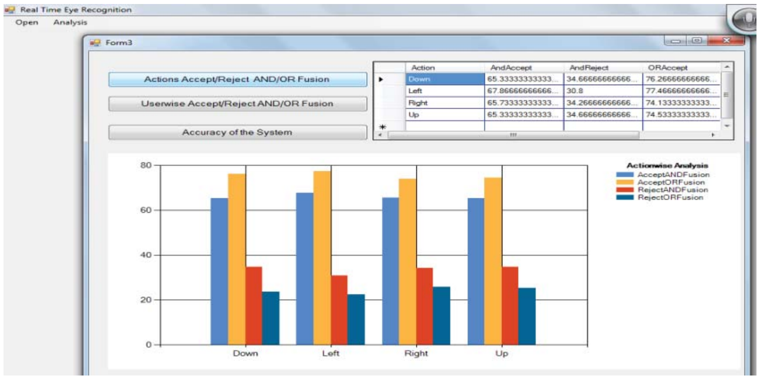
  Step3. A pair of fused feature vectors, Zi and Zj , are then compared with the threshold (stad). If stad > t (and 0, otherwise), t is a pre-specified threshold, and k is the dimensionality of the fused feature vector.

  n_MM_simple_sumRule(u) = n_iris_MM + n_Speech_MM;

- Quality Based Score Fusion: Quality based score fusion algorithm [7] involves quality metric of both modalities, Normalisation of scores is done . Matching scores from matching  along with the quality information are forwarded to fusion process.

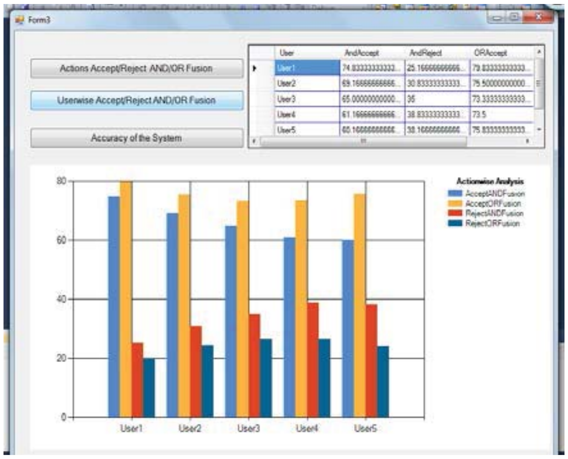
## V. EXPERIMENT AND RESULT

- AND-OR FUSION RESULTS:

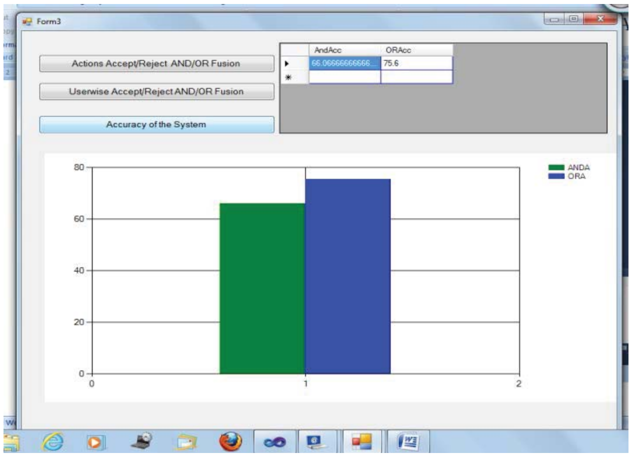


| Action | AndAccept | AndReject | ORAccept |
|---|---|---|---|
| Down | 65.33333333333... | 34.66666666666... | 76.26666666666... |
| Left | 67.86666666666... | 30.8 | 77.46666666666... |
| Right | 65.73333333333... | 34.26666666666... | 74.13333333333... |
| Up | 65.33333333333... | 34.66666666666... | 74.53333333333... |

Figure 1.  Graph Showing Actionwise Accept/Reject Analysis



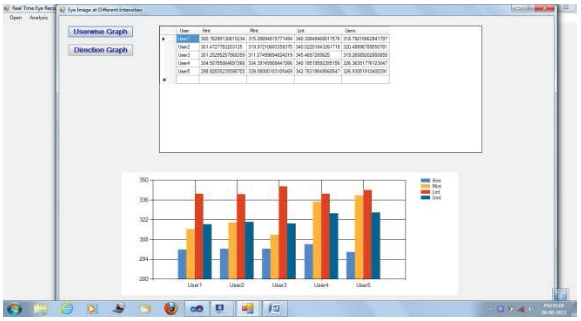
| | User | AndAccept | AndReject | ORAccept |
|---|---|---|---|---|
| ▸ | User1 | 74.83333333333... | 25.16666666666... | 79.83333333333... |
| | User2 | 69.16666666666... | 30.83333333333... | 75.50000000000... |
| | User3 | 65.00000000000... | 35 | 73.33333333333... |
| | User4 | 61.16666666666... | 38.83333333333... | 73.5 |
| | User5 | 65.16666666666... | 38.16666666666... | 75.83333333333... |


Form3
Actions Accept/Reject AND/OR Fusion
Userwise Accept/Reject AND/OR Fusion
Accuracy of the System
ANDA
ORA
80
60
40
20
0
0
1
2

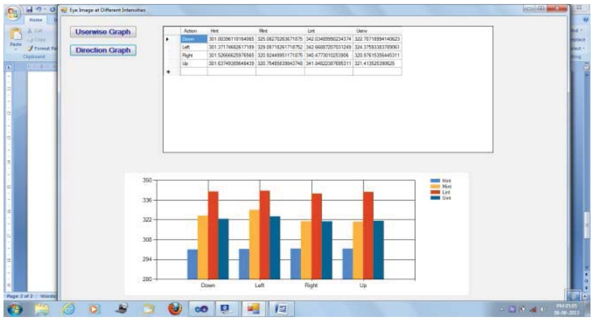
| | AndAcc | ORAcc |
|---|---|---|
| ▸ | 65.06666666666... | 75.6 |
| * | | |

Figure 2. Userwise Accept/Reject Analysis

Figure 3. AND/OR Fusion wise Accept/Reject Analysis



| | User | Hint | Mint | Lint | Uenv |
|---|---|---|---|---|---|
| ▸ | User1 | 300.76290130615234 | 315.28694915771484 | 340.32648468017578 | 318.79210662841797 |
| | User2 | 301.4727783203125 | 319.97210693359375 | 340.02201843261719 | 320.48896789550781 |
| | User3 | 301.25295257568359 | 311.07499694824219 | 345.4687265625 | 319.26589202880859 |
| | User4 | 304.50785064697266 | 334.39749908447266 | 340.18519592285156 | 326.36351776123047 |
| | User5 | 298.92835235595703 | 339.08000183105469 | 342.78319549560547 | 326.93051910400391 |
| * | | | | | |



| | Action | Hint | Mint | Lint | Uenv |
|---|---|---|---|---|---|
| ▸ | Down | 301.00396118164065 | 325.08270263671875 | 342.03489990234374 | 322.70718994140623 |
| | Left | 301.37174682617189 | 329.08718261718752 | 342.66887207031249 | 324.37593383789061 |
| | Right | 301.52666625976565 | 320.92449951171875 | 340.4773010253906 | 320.97615356445311 |
| | Up | 301.63749389648439 | 320.75485839843748 | 341.84822387695311 | 321.413525390625 |
| * | | | | | |

Figure 4.User wise Analysis on the basis of different Environment

(High, Low, Medium, Constant Intensity)

Figure 5. Direction wise Analysis on the basis of different Environment:

(High, Low, Medium, Constant Intensity)

- Sum Rule-based fusion and Quality score based Fusion:

The Sum Rule-based fusion and Quality score based Fusion are processed by the set of users on the respective application, with different set of conditions. The performance of each user for different parameter like genuine recognition rate for speech and eye (SPEECH and EYE), processing time required by the user, false acceptance rate for eye and speech (FAREye and FARSpeech) and correct fusion recognition rate is computed .The other part is analysis of error which occurs during human computer interaction, when fusion of multiple inputs processed together causes ambiguities, insufficient language understanding while recognizing speech, mutual disinfluencies of multiple input while giving command to application, unsynchronized inputs from the user i.e. big pause between speech and gesture, etc.., considering these error problems, it is necessary to perform error analysis and compute the error rate.

The Respective graphs shows the performance of application with respect to different parameter like processing time, complexity, accuracy, error rate, etc.., in various condition as ideal set of condition and un favorable conditions for sum based fusion and Quality score based fusion of eye and speech are as shown below:

The performance of various users for sum based fusion with respect to processing time required for complete execution is analyzed in following graph shown in fig.

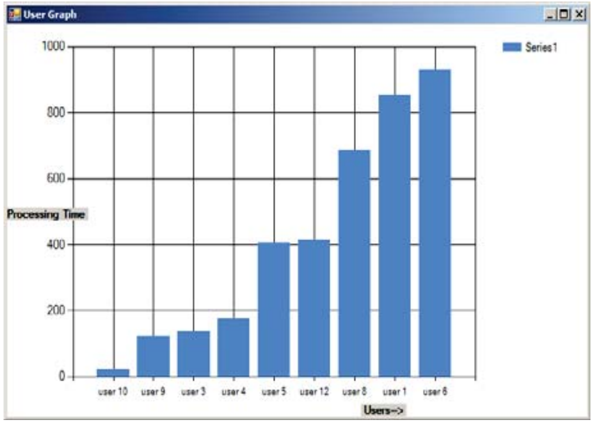


Figure 6.Graph showing the performance of various users for sum based fusion in uniform condition

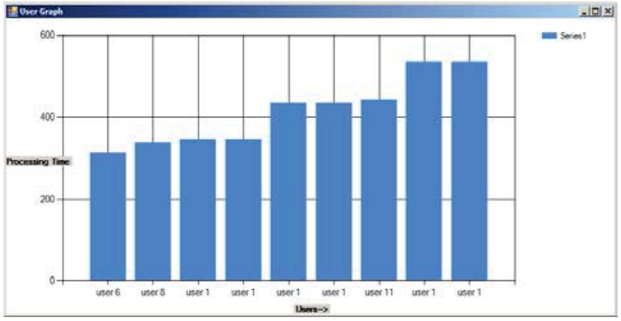


Figure 7. Graph showing the performance of various users for Quality score-based fusion in uniform condition

The various techniques are evaluated on the basis of their performance which consists of time complexity i.e., time required by the technique to complete one iteration for single set of input. The fig shows the graph which compares sum based and quality score based fusion on time complexity parameter.

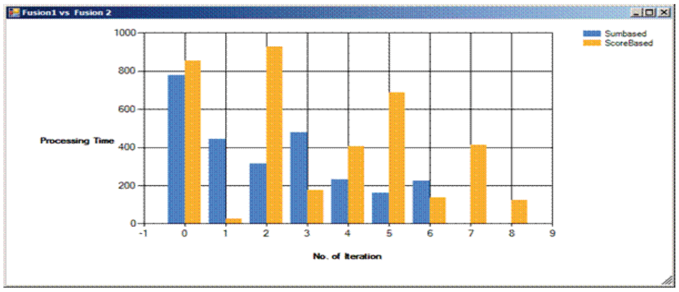


Figure 8. Graph showing the performance of sum based and score based fusion with respect to processing time vs. number of Iteration in uniform condition

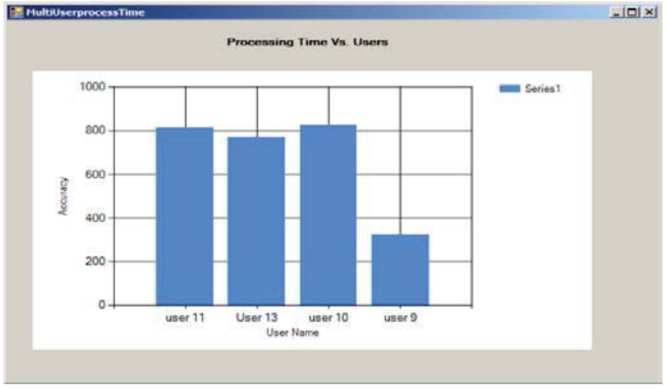


Figure 9. Graph showing the performance of various users for sum based fusion with respect to Processing Time in Non-Uniform Condition.

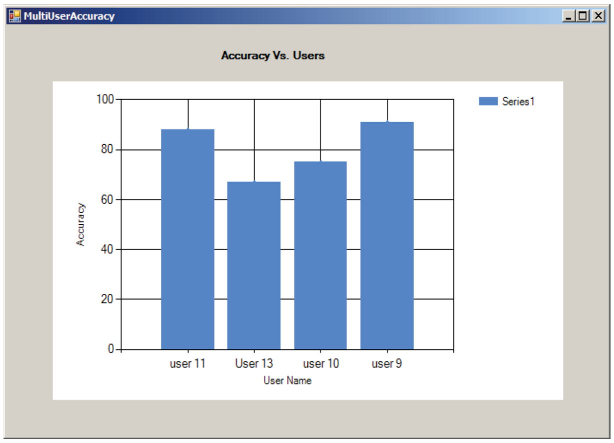


Figure 10. Graph showing the performance of various users for sum based fusion with respect to Accuracy in Non-Uniform Condition

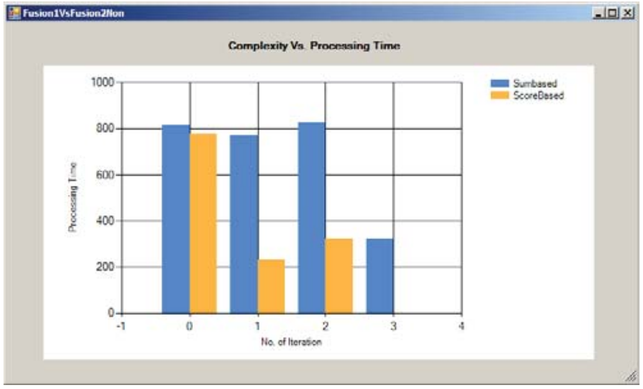
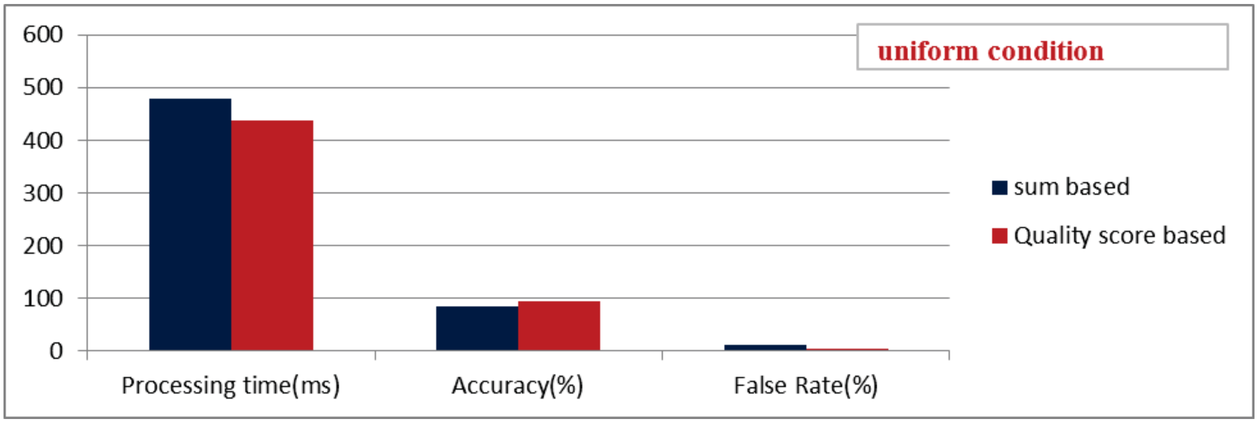


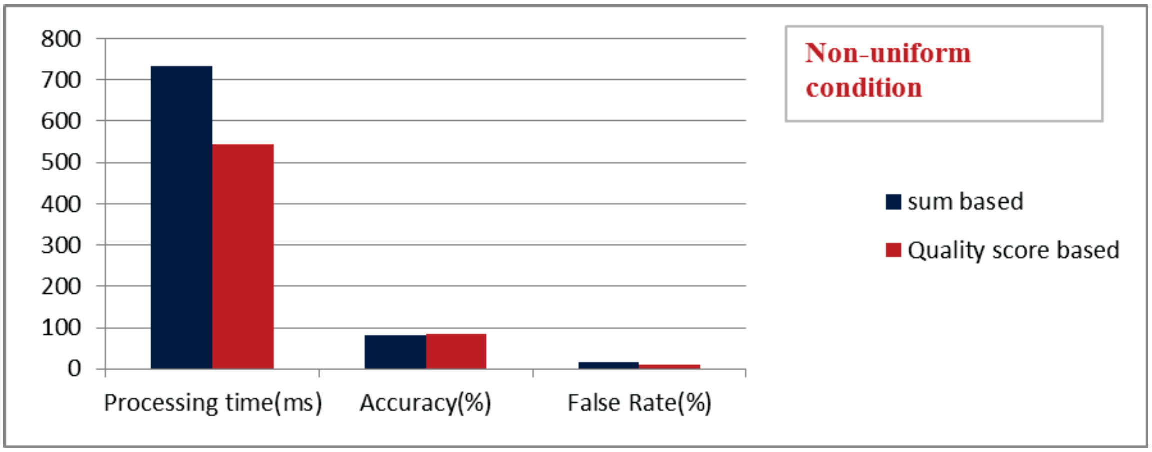
Figure 11. Graph showing the performance of sum based and score based fusion with respect to processing time vs. number

of Iteration Non-Uniform Condition

Figure 12. Comparative Results :

Graph





## VI.CONCLUSION

This work compares the fusion techniques like AND/OR ,Sum Rule-based Fusion and Quality Score based Fusion and evaluate the optimised fusion technique with respect to the parameters like processing time, accuracy, error rate. With Reference to the results statistics and graph analysis in both uniform and non-uniform condition, it is concluded that Quality score based fusion is optimised and error-free fusion technique.

## REFERENCES

[1] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh and Mo Nours Arab, (2008), "Human-Computer Interaction: Overview on State of the Art", International journal on smart sensing and intelligent systems,ol. 1, no.1, March 2008.

[2] Matthew Turk , (2013) "Multimodal interaction: A review", doi /10.1016/ 2013.07.003 Pattern Recognition Elsevier journal 2013.

[3] Joyce Y. Chai Zahar Prasov, Pengyu Hong, "Performance Evaluation and Error Analysis for Multimodal Reference Resolution in a Conversation System".

[4] 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.

[5] Meriem Bendris, Delphine Charlet, "Introduction of Quality Measures In Audio- Visual Identity Verification", IEEE 978-4244 2354/2009

[6] Hanaa S. Ali, Mahmoud I. Abdalla, "Score-Level Fusion for Efficient Multimodal Person Identification using Face and Speech", (IJCSIS) International Journal of Computer Science and Information Security, Vol.9, No. 4, April 2011.

[7] Norman Poh, Thirimachos Bourlai, Josef Kittler, Lorene Allano, Fernando Alonso-Fernandez, Onkar Ambekar, John Baker, Bernadette Dorizzi, Omolara Fatukasi, Julian Fierrez, Harald Ganster, Javier Ortega- Garcia, Donald Maurer, Albert Ali Salah, Tobias Scheidat, and Claus Vielhauer, "Benchmarking Quality- Dependent and Cost-Sensitive Score-Level Multimodal Biometric Fusion Algorithms", IEEE transactions on information forensics and security, vol. 4, no. 4, December 2009.