Review of Optimised Fusion Techniques in Human Computer Interaction

[1]Madhuri R. Dubey, [2] S. A. Chhabria, [3]R.V. Dharaskar

[1]Student, M.Tech CSE, [2]HOD, IT Department, [3]Director,
[1,2]G.H.Raisoni College of Engineering, Nagpur, [3]Matoshri Pratishthan's Group of Institutions, Nanded
Email : [1]madhuridubey8@gmail.com, [2]sharda.chhabria@raisoni.net, [3]rvdharaskar@yahoo.com

Abstract—This paper provides an overview of various optimisation techniques available in multimodal fusion in human-computer interaction. Humans use a variety of modes of information to recognize people and understand their activity. Fusion of multiple sources of information is a mechanism to robustly recognize human activity. The overview includes the basic definitions and approaches of quality based fusion and contains the survey of various technique evolved for fusion with optimisation. This paper also highlights the features of quality based fusion over traditional fusion. It proposes the optimized fusion of various modes of senses and error analysis in human-computer interaction applications.

Index Terms— Multimodal fusion, Human-Computer Interaction, Quality based Fusion, Optimisation Technique.

I. INTRODUCTION

Human-computer interaction is a branch which focus on the design, evaluation and implementation of interactive computing systems for human use and with the analysis of their surrounding environment. A future goal of human computer interaction is to design systems that minimize the barrier between the human's cognitive model of what they want to achieve and the computer's understanding of the user's task. The advent of widespread innovative technology tends to shift the human-computer interaction paradigm from a traditional system based usage to a more natural and reliable manner in which users may interact with the system.

Speech and eye are important modality in human-human and human–computer interactions. Speech signals provide valuable information required to understand human activities and interactions. Speech is also a natural mode of communication for humans. Human activity in a scene is usually monitored using arrays of audio and visual sensors like camera, microphone. [1]

The combination of input from various modes of senses enables the development of human intelligent systems known as fusion of multiple modalities. 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. The fusion of multiple modalities can provide alternative information and increase the accuracy of the overall decision making process.

The fusion schemes are categorised into the following three methods:

- Rule-based methods :

The rule-based fusion method includes a variety of basic rules of combining multimodal information. These include statistical rule-based methods such as linear weighted fusion (sum and product), MAX, MIN, AND, OR, majority voting.

- Classification based methods :

This category of methods includes a range of classification techniques that have been used to classify the multimodal observation into one of the pre-defined classes. Such as Support vector machine, Bayesian inference network, Dempster–Shafer theory, Dynamic Bayesian networks, neural networks and Maximum entropy model, etc. [15]

- Estimation-based methods :

The estimation category includes the Kalman filter; extended Kalman filter and particle filter fusion methods. These methods have been primarily used to better estimate the state of a moving object based on multimodal data. For example, object tracking, multiple modalities such as audio and video are fused to estimate the position of the object. [2]

The fusion of different signals can be performed at different levels as shown below; [2][9]

1) Raw data or the feature level,
2) Score level
3) Decision level

The raw data or feature level fusion can be compatible for all modalities and a common matching algorithm used. If these conditions are met, the separate feature vectors of the modalities are easily fused into a single new vector.

The fusion at score level is evaluated by calculating a similarity or dissimilarity (distance) score for each single modality.

The modality results in its own decision are referred as Decision Level Fusion; in case of a verification scenario this is a set of true and false.

In short, the voting (majority decision) or a logical "AND" or logical "OR" decision can be computed. Decision level fusion is known for easiness and the guaranteed availability of all single modality decision results.

II. OPTIMIZATION IN FUSION

A. What is Optimised System?

The definition of optimization is "The process which used to make a system or design as effective or functional as possible".

The system which maximize the performance and minimize recognition error subject to maximum permissible costs can be called as Optimised system. And to achieve this state, we specify constrained optimization problem. [7]

The optimum system can be defined in terms of the reduction in ambiguity or error in the estimation process. The process of estimation should be more reliable and accurate if the ambiguity or error in the underlying estimation can be minimized. For the estimation process, when fusion is processed the analysis of only quality metric is evaluated to minimize number of parameter for fast and quality performance. [11]

Although in current state, one cannot minimize both the error and the cost simultaneously. To solve constrained optimization problem the system needs repeated evaluation of the fusion module performance. [7]

It is important that to quantify performance so that systems can be optimized within real time application. It is also important to state these performance measures up front which are as follows;

- Time Responsiveness
- Throughput
- Reliability and availability

To end idle sitting of user in front of system, performance is usually equated with response time.

In short, performance measures in throughput, which is measured in terms of the number of transactions processed in unit interval of time.

A system that is not functioning has zero performance. An optimized system must be extremely reliable.

B. Optimisation Techniques for fusion

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 previously.

There are various optimization techniques available in multimodal fusion proposed in the set of papers, some are reviewed below;

i) Reference resolution Technique is used to find the most proper referents to referring expressions. This technique focussed on graph matching algorithm. It uses two graph i.e. referring graph that captures referring expressions from speech utterances as input. The node leads to referring expression, consists of the semantic information extracted from the expression, also the edge represents the semantic and temporal relation between two referring expressions. Referent graph that represents all potential referents like objects selected by the gesture, objects in the conversation history, and objects in the visual focus, etc.. node and edges contain similar information as that of referring graph.

- Algorithm Steps:

Step 1: Construct word confusion network:

Align words from n-best list into equivalence classes.

1.1: clustered the starting and end timestamp of various instances of same words.

1.2: equivalent class with common time ranges are merged

1.3: compute probability of all utterances hypothesis containing this word and the probability is assigned using rank list of merged clustered.

Step 2: Extract referring expression from word confusion network (WCN):

2.1: Applied set of grammar rules to parse the confusion network of utterances.

2.2: Identified sub phrases of speech utterance in WCN.

Step 3: Resolve Referring expression:

Assume r_j referring expression is resolved to the top 'k' potential referent object according to probability $P(o_i/r_j)$.

$P(o_i/r_j)= (AS(oi)^{\alpha} \times Compat(oi, rj)^{1-\alpha})/(\sum_i AS(oi)^{\alpha} \times Compat(oi, rj)^{1-\alpha})$

Where;

AS: Attention silence score for object o_i

Compat: compatibility score which specify object o_i is compatible with referring expression r_j.

α: importance weight of range [0----1], initially α=0.0

Step 4: post-prune:

The resulting set of (referring expression, referent object set) pairs is pruned to remove pairs which consist of :

4.1: the pair has a confidence score equal to or below a predefined.

4.2: the pair temporally overlaps with a higher confidence pair.

Step 5: stop processing when the expression is resolved.

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. [12]

ii) The Chernoff measures method provides an upper bound of the minimum Bayes error given p(y|k). This technique is well suited for a two class problem when p(y|k) is a multivariate normal distribution for each k [2], with mean μ^k and covariance $\sum^k$. The classification error can be bounded can be referred as Chernoff bound. This bound used to derive an upper bound for Half Total Error Rate (HTER) which is defined as the average of False Acceptance Rate (FAR) and False Rejection Rate (FRR), both of which measure how probable a system accepts an impostor claim and rejects a genuine claim, respectively. [7]

iii) Genetic algorithm (GA) and heuristic algorithm like stimulated annealing and evolution strategy (ES), etc.. are used as optimised algorithm. It states that the evolution range is updated during the optimization process by evaluating the convergence rate. The "self-adaptation" is the unique feature of the ES which involves mutation, crossover, shaking.

- Algorithm Steps:

Step 1: Initialization of $\alpha_{min,i}$, $\alpha_{max,i}$, and $\alpha_{init,i}$ for design variable where;

α_i : Evolution range for i^{th} design variable. child generation is generated P_i within $[P_i- \alpha_i, P_i-+\alpha_i]$ when $\alpha_{min,i} < \alpha_i < \alpha_{max,i.}$

$\alpha_{min,,i,}$: Minimum distance between elite solution

$\alpha_{max,i}$: Maximum distance between elite solution

$\alpha_{init,i}$:Initial value for $\alpha_{i.}$

Step 2: Generation of $\alpha_{init,i}$:

Find initial population of elite set.

Size of initial population =λ*μ; where μ is selected using stimulated annealing approach i.e. finding best of the solution among other.

Step 3: Generating children and restricted evolution

Create λ new children within mutation range $[P_i- \alpha_i, P_i-+\alpha_i]$ the restricted solution finds local optimum solution for each elite member.

Step 4: Mutation:

If P_i is more improved than α_i then replace P_i by $\alpha_{i.}$

If ∀ P_i contains α_i and objective function of P_i has worse value than other solution then removed value from elite set ξ.

Step 5: Shaking:

The ξ +P solutions are randomly generated in the whole search space outside the mutation ranges of existing elite solutions called as number of shaking solutions.

Step 6: Annealing:

The ξ removed solutions are replaced by the new solutions generated by the shaking process.

Step 7: Convergence Check:

Repeat step 3-6 until solutions are more improved i.e. optimised.

This approach prevents solutions from clustering with their neighbours and allows only one solution to survive at each level. Hence, this is more efficient and practical than the conventional approaches. [8]

iv) The Kalman filter (KF) allows for real-time processing of dynamic low-level data and provides state estimates of the system from the fused data with some statistical significance.

This filter is used in a linear dynamic system model with Gaussian noise to be assumed. KF does not require preserving the history of observation and only depends on the state estimation data from the previous timestamp.

- Algorithm Steps:

Step 1: Assume: X = Predicted, P = Covariance, Y = Residual, M = Measured, S = Residual Covariance, R = Minimal innovative covariance, K = Kalman gain, Q = Minimal update covariance of P, H = Rolls actual to predicted, I = Identity matrix.

Step 2: Measured input and subdivide into frames and features w.r.to time.

Step 3: Move prediction from old time to new time to detect motion of frames.

$$X = F*X + B*U$$

$$P = F*X*F^T + Q$$

Step 4: Update: Residual, Residual Covariance, Kalman gain, Covariance and Predicted frameset. Process step 2, 3 again

Step 5: Detection of illumination and apply filter to the current time frame

5.1: Initialize P always to a diagonal large matrix.

5.2: State prediction covariance

$$P(k+1|k) = F(k)P(k|k)F(k)'+Q(k)$$

5.3: Measurement prediction covariance:

$$S(k+1) = H(k+1)P(k+1|k)H(k+1)'+R(k+1)$$

5.4: Filter Gain

$$W(k+1) = P(k+1|k)H(k+1)' S(k+1)^{-1}$$

5.5: Updated state covariance:

$$P(k+1|k+1) = P(k+1|k) - W(k+1)S(k+1)W(k+1)'$$

v) Optimal coupling Method 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. At the stage of fusion,

1. Selects an optimal pair of cutpoints in the audio track, based on the minimization of the auditory join cost.
2. Select the cutpoints in the visual mode so the video clusters can be fused together.
3. Two different approaches were implemented and evaluated as shown ;

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.

However, even the smallest difference between the cutpoints in both modes causes a discrepancy between the length of the audio track and the length of the video track of the selected multimodal segment.

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.

- Algorithm Steps:

Step 1: For each system i = 1, ---- , N:

- Calculate the optimal λ for yi using (2)
- Transform yi using (1):

$$y_i\text{norm} = T(yi, \lambda *)$$

Step 2: Let $y = [y_1\text{norm}, \text{-----}, y_N\text{norm}]'$ where yi's is the transformed match score

Step 3: Compute $p(y|k) = N(y|\mu^k, \_\sum^k)$

Step 4: For each combination $y_c \in P(\{yi|\forall i\}) - \emptyset$ (indexed by c):

• Calculate $\text{criterion}_c = 1/2 \min_\beta \exp(-k(\beta|\mu_c^k, \sum^k_c, \forall k)))$

Step 5: Output: $\arg \text{sort}_c \{\text{criterion}(c)\}$

This optimal coupling algorithm has three highlighted parameters: [5-6]

- maximal local audio lead (negative desync),
- maximal local video lead (positive desync) and
- search length parameter

vi) Sequential Forward Floating Search (SFFS) with support vector machine as wrapper to employ classification error as optimization criterion and avoid NP-hard exhaustive search explained in (Schuller et al., 2005). It can be similar to optimized method rather than finding single attributes of high relevance for the system. Here audio video fusion is considered features in one pass to point out key features of audio and video. The optimal number of features is determined on the basis of highest accuracy in between them throughout selection process.

This method saves computation time considering real-time processing and boosts performance as some classifiers are susceptible to high dimensionality. [13]

vii) Hidden Markov models are widely used in science, engineering also in various areas like speech recognition, optical character recognition, machine translation, bioinformatics, computer vision, etc.

The Hidden Markov Model (HMM) is a variant of a finite state machine having a set of hidden states Q, an output alphabet (observations) O, transition probabilities A, output (emission) probabilities B and initial state probabilities Π.

HMM is said to be a triplet of (A, B, Π).

- Algorithm Steps:

Assume $\alpha_t(i)$ be the probability of the partial observation sequence $O_t = \{o(1), o(2), \ldots, o(t)\}$ to be

produced by all possible state sequences that end at the i-th state.

Step 1: The probabilities for the single-symbol sequence are calculated as a product of initial i-th state probability and emission probability of the given symbol o(1) in the i-th state.

$$\alpha_1(i) = p_i\, b_i(o(1)) \text{ , } i = 1, \ldots, N$$

Step 2: The recursive formula is applied to calculate $\alpha_t(i)$ for some t.

Step 3: To calculate $\alpha_{t+1}(j)$

3.1 multiply every $\alpha_t(i)$ by the corresponding transition probability from the i-th state to the j-th state

3.2 sum the products of all states

3.3 Multiply the result by the emission probability of the symbol o(t+1).

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j)\, a_{ji}\right] b_i(o(t+1))$$

Where, i =1, ... , N , t =1, ... , T - 1

Step 4: Iterating the process to calculate $\alpha_T(i)$, and make summation of all states.

$$P(o(1)o(2)\ldots o(T)) = \sum_{j=1}^{N} \alpha_T(j)$$

Step 5: Stop processing when the desired output is achieved.

C. What is non Optimized system?

In Non- optimization, the system's goal is only to reduce the cost of compilation and to make debugging and generate the expected results. Statements are independent: if the program get stop with a breakpoint between statements, the assignment of new value to any variable take place and user get exactly the results which is expected from the source code.

Whereas, optimization makes the system to attempt the improved performance and the expense of compilation time and possibly the ability to debug the program will be minimized.

III. QUALITY BASED FUSION

Quality-dependent fusion algorithms aim to dynamically combine several classifier outputs as a function of automatically derived sample quality.

- Quality-dependent evaluation and
- Cost-sensitive evaluation.

i) Quality-dependent evaluation:

It involved client-specific or user-dependent fusion where one can train a fusion classifier that is tailored to each identity claim.

Quality measures are expected to provide measurements designed to capture changes in ways that could usefully be exploited in the fusion process.

ii) Cost-sensitive Evaluation:

It concerned with handling missing information. Whenever, one or more subsystems are not operational due to failure-to-acquire or failure-to-match a biometric sample, the fusion system to be able to output a combined score.

In a cost-sensitive evaluation scheme, one considers a fusion task as an optimization problem whose goal is to achieve the highest performance (as a function of false acceptance and false rejection decisions) at a desired minimum cost. [3]

Approaches of Quality based fusion:

Quality-based fusion has following two approaches depending on the role of quality measures:

- Feature-based approach:

Feature-based fusion classifier treats quality measures as another set of features, like the expert outputs (scores). Classifiers in this category typically concatenate the expert outputs and quality measures into a single vector.

It uses quality measures directly as features.

- Cluster-based approach:

Cluster based approach, first clusters the quality measures into a number of clusters. Then, for each cluster, a fusion strategy is designed. This cluster-based approach can be seen as a divide-and-conquer strategy. It breaks the fusion problem into multiple, smaller but also simpler ones. [4]

IV. PROPOSED WORK

A wide variety of fusion techniques applied for various application of Human computer interaction but the technique gives the optimized result in terms of speed, time, accuracy will be needed at current scenario.

- Find the optimized fusion technique out of the bulk of available techniques
- Analyzed the error at every level of fusion to make error free system for any application.

➢ Objectives

To evaluate performance of various fusion techniques and detect the optimized Fusion technique which will be more efficient in terms of accuracy and results. And analyzes error occur during human machine interaction when fusion of multiple inputs processed together, small changes in input quality or environment caused ambiguities. Also, find new and optimized Interaction method in the field of Human Computer Interaction.

➢ System Overview

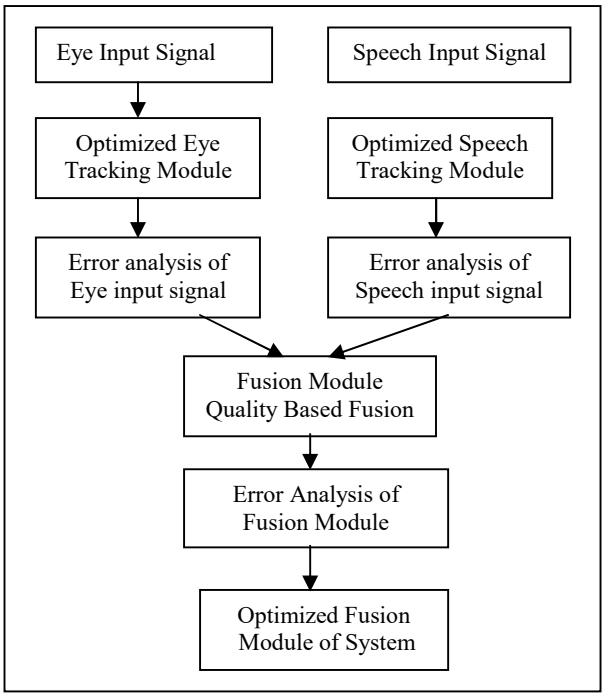

Figure 1. Overview of System

V. SCOPE

- The better real time performance is depend on the accurate optimized fusion model considering parameter like accuracy, speed, complexity, etc..
- In human computer interaction, fusion is used to identify multiple commands or input interaction of human like speech, gestures, etc.
- The optimized and error free output of fusion of multiple modes will meet the timeliness by reducing re-computation of inputs.
- It can address and assist disabled people (as persons with hands disabilities), which need other kinds of interfaces than ordinary people. In such systems, disabled users can perform work on the PC by interacting with the machine using voice and head movements, etc. [14]

VI. CONCLUSION

This paper highlights various approaches for multimodal human-computer interaction. Also it discusses techniques for fusion of human modes of senses like eye, speech, hand, etc. the optimised system and the need of optimisation in multimodal fusion, and a variety of emerging techniques and approaches of quality based fusion.

REFERENCES

[1] 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.

[2] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, Mohan S. Kankanhalli, " Multimodal fusion for multimedia analysis: a survey", Multimedia Systems, Springer- 2010.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Norman Poh and Josef Kittler, "On Using Error Bounds to Optimize Cost-Sensitive Multimodal Biometric Authentication"

[8] 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.

[9] 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.

[10] Satoshi Tamura, Koji Iwano and Sadaoki Furui, "Toward robust multimodal speech recognition".

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.