Multimodal Interaction in Modern Automobiles

Ashish Khare
Hiranmay Ghosh
Sujal Subhash Wattamwar
TCS Innovation Labs Delhi
249 D&E Udyog Vihar Phase 4
Gurgaon Haryana India
{ashish16.k | hiranmay.ghosh | sujal.wattamwar}@tcs.com

Aniruddha Sinha
Brojeshwar Bhowmick
K S Chidanand Kumar
TCS Innovation Labs Kolkata
BIPL Bldg Salt Lake Electronics Complex Kolkata West Bengal India
{ aniruddha.s | b.bhowmick | kschidanand.kumar}@tcs.com

Sunil Kumar Kopparapu
TCS Innovation Labs Mumbai
ODC G SDC-V Yantra Park
Subhash Nagar Pokharan Road 2,
Thane(West) Maharashtra India
sunilkumar.kopparapu@tcs.com

ABSTRACT
This paper describes a few innovative solutions for application of multimodal interaction techniques in modern automobiles to ensure driving comfort, safety and security. The solutions are based on computer vision and speech processing techniques.

Keywords
Multimodal interaction, hand gesture recognition, speaker verification, driver fatigue detection.

INTRODUCTION

Modern cars have gone beyond providing a means for commutation and tend to create a personal space for its inhabitants. The provision of advanced entertainment systems, climate control equipment, and navigation aids in modern cars are just some of the examples of such transition. Provision of such additional equipments in a car brings in some new challenges. Control of secondary equipment while the driver needs to concentrate on driving can cause distractions and be a potential safety threat [1, 2]. Long and lonely drives also contribute to driver fatigue resulting in fatal accidents [3, 4]. Security of the vehicles in urban society has become another major issue in the recent times [5].

In this context, researchers are seeking multimodal interaction techniques with the automobile for enhancing driving comfort, safety and security. Multimodal interaction refers to use of several natural modes of communication, such as gesture, gaze and speech, to complement traditional electro-mechanical interaction devices such as control buttons, joysticks and specially designed levers. Moreover, several bio-metric techniques can be used for vehicle security and enhancing driving safety by authenticating the driver and ascertaining his physiological condition. Several centers of TCS Innovation Labs have been working together with leading automobile manufacturers to provide such multimodal interfaces. In this paper, we provide a few examples of innovative multimodal interfaces in an automobile to provide driving comfort, safety and security.

MULTIMODAL CONTROL OF IN-VEHICLE EQUIPMENTS

The secondary equipments in an automobile, e.g. entertainment system, climate controls and other navigational aids are traditionally controlled with buttons, touch-screens and remote devices by the driver. Operating a plethora of control buttons of the various equipments while driving is not only inconvenient but can cause distraction to the driver causing the automobile going out of control and resulting in serious accidents. 'Remote' devices and placement of the buttons at vantage points like embedded on the steering wheel partially solves the problem. We propose interaction with such equipment with gesture and speech which requires little distraction. The motivation behind using gesture and speech together is to improve the robustness of the system and to provide alternative modes of communication. While gesture requires short diversions of visual attention, speech recognition may not be robust in noisy driving environment.

Figure 1 depicts the system architecture. The car is equipped with a camera and a microphone to pick up the drivers voice and hand gesture. The driver of the car has two options to control an in-vehicle system, say the music player. He could either speak a word from the vocabulary (start, stop, eject, etc) to control the car audio system or alternatively he could create a gesture (using a predefined gesture vocabulary) with his hand. Speech and gesture recognition technologies are used to interpret the spoken words and the gesture made. The data obtained from the two channels are fused in context of the previous user interactions and the current instrument status. In absence of tactile or visual feedback, the interpreted action request is spoken out to the user.

Gesture recognition systems responds either on signal emitted by some implants worn on the human hand or on computer vision techniques to track naked human hand [6]. While the former approach is more reliable, it is inconvenient to wear the implants while driving.

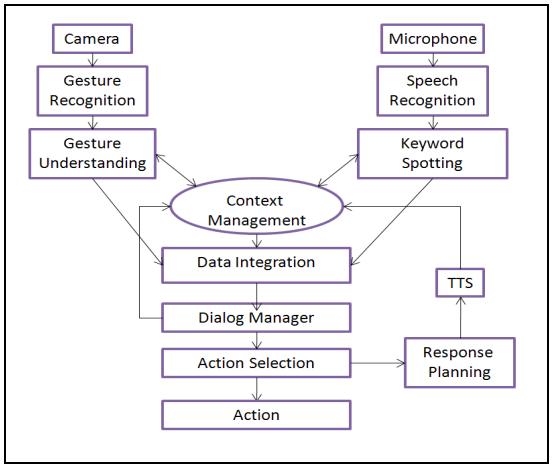

Figure 1. Multimodal interface for in-vehicle instrument control.

This motivates us to use computer vision techniques for gesture recognition for control of in-vehicle systems. Further, studies indicate that visual attention is the most important requirement for the driver. So, we have designed the gesture vocabulary to contain a few static gestures, which require short or no distracting glances form the primary driving tasks. Our gesture vocabulary include a few directional gestures (e.g. up, down, forward and backward) and outstretched fingers indicating numbers 1-5. The closed palm forms another distinct gesture. The directional gestures can be interpreted as volume up/down and skip forward/backward and the numbers to indicate a particular track for a music system. Figure 2 provides a block diagram of the gesture recognition system.

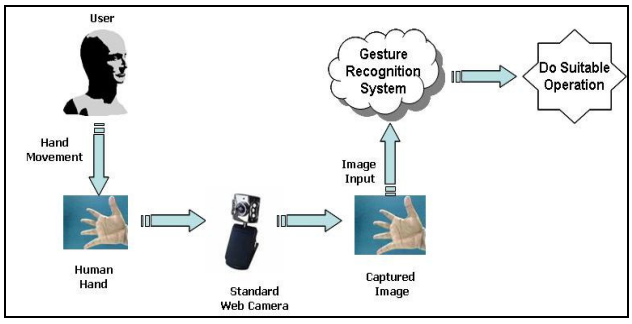

Figure 2. A Typical Gesture Recognition System.

We have used the algorithm as described in [7] for counting outstretched fingers and a Bayesian Network based machine learning approach for recognizing the directional gestures. While the former algorithm is restricted to counting fingers only, the latter approach can be used for detecting any arbitrary set of gestures. In this approach, the training set comprising a large number of labeled gestures are clustered based on some extracted media features, e.g. edge-histogram and template. A naïve Bayesian Network where the root node represents the gesture states and the leaf nodes represent the feature states (clusters) is trained using these labeled gestures. The features of a test gesture are extracted and the leaf nodes of the Bayesian Network are instantiated accordingly. The gesture is recognized on the basis of posterior probability of the gesture states in the root node as a result of belief propagation in the Bayesian Network.

The speech recognition system that complements gesture recognition in controlling in-vehicle instruments has a small vocabulary (words associated with audio control) isolated word recognition (IWR) speech engine built onboard the car computer to recognize the voice command and take appropriate action. The IWR engine is based on dynamic time warping (instead of the regular HMM's because the DTW system is a small footprint system which can be easily embedded onto the computer on the vehicle itself) to recognize voice commands. The command and control system wakes up when a particular word or a sequence of words (pre configured; called the *wake up* word) is spoken and expects the user to speak (following the *wake up* word) a word or a sequence of command words. The onboard speech recognition engine after necessary pre-processing (segmenting the speech signal into frames; filtering) extracts feature which are used to match the spoken command with the vocabulary words; once the command is recognized a signal is sent to the control mechanism onboard the car and the command is executed (Eg. "volume up", "channel three" etc). The speech recognition engine though has a small vocabulary, the system offers some flexibility to use different phrases to execute the same operation, e.g. "wiper on" or "switch on wipers" have the same connotation.

ACCESS CONTROL OF VEHICLES

Gaining keyless access to one's own car using biometrics is a feature that is gaining popularity and has been implemented in concept and high level entry cars. But the state of the art in speech biometric especially as an embedded solution and the environmental conditions that the owner of the car might be in (highway, noisy street, and garage) present challenges to the speech based verification system.

We propose a multi authentication speaker verification system to enable the owner of the car gain access to his vehicle with small false rejection ratio and small false acceptance ratio. The idea is to have a small footprint embedded speaker verification system installed inside the car and a robust speaker verification system (more accurate) installed on a server.

When wanting to gain access to the car, the car owner speaks simultaneously into his mobile phone and to the microphone mounted discretely on the car. The same

speech signal is authenticated by the *less* powerful speaker verification system inside the car and simultaneously the speech via the mobile phone is delivered to the speaker verification server for verification. Only when both the speaker verification system on board the car and the system on the server verify the identity of the user is the user allowed access to the car.

The high level architecture of the system is depicted in Figure 3. The user who wishes to gain access to the car speaks the pass phrase simultaneously into the mobile phone (1 in Figure 3) and the car (2 in Figure 3). An embedded speaker verification system in the car verifies the identity and flags it as verified (Yes) or not verified (No) (5 in Figure 3); simultaneously via 1, 3 the speaker is verified at the server and a verified or not is marked (4 in Figure 3). Both the verifications are combined and a decision to given access to the car is invoked if both server and the embedded verification system return a Yes (6 in Figure 3). The server verification path is shown in red while the embedded verification path is shown in blue in Figure 3.

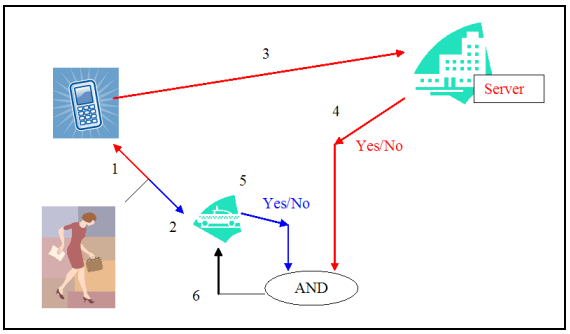

Figure 3. High level architecture for speech enabled access to vehicles

The speaker verification system onboard the car is based on using dynamic time warping (DTW) and cepstral LPC (CC-LPC) coefficients which allows for a small footprint verification system while the server based system is based on the more reliable Mel Frequency Cepstral Coefficients (MFCC) and uses hidden Markov models (HMM) to verify the user. This configuration of multi parameter, multi path, multi model architecture leads to a robust verification system as against using a speaker verification system onboard the car alone.

Speaker verification is the process of verifying the claimed identity of an individual using his voice characteristics (popularly called voice print). The speaker verification task is different from the process of speaker identification. Speaker verification tests the hypothesis that a certain individual X is the speaker of a given utterance (``Are you who you claim to be?'', namely, a test with two possible outcomes – yes or no), whereas a speaker identification system determines if the speaker of a given utterance is among a set of 'N' registered speakers or is an unregistered speaker (``Do I know you?'',' n+1' possible outcomes). This conceptual difference also causes differences in accuracy, execution time, scalability and applicability between the verification and the identification process. It is not difficult to notice that for identification process both accuracy and execution time critically depend on the size of the set of registered speakers: the larger this set size, the lower accuracy and longer the execution time; while for the verification process there is no significant affect. In our scenario, speaker verification makes sense assuming that the car is owned by an individual.

Speech recognition in general and speaker verification in particular use typically extract two sets of parameters; one set is based on the modeling of speech production system as a linear finite impulse response (FIR) filter and popularly called the LPC parameters, while the second set is based on the modeling of the auditory perception of speech which leads to the MFCC parameters. In our approach we use both these parameters extracted from the spoken speech which gives the overall system the required robustness to perform the verification more efficiently.

Further in speech literature there are two different methods of comparing speech signals (a) one is a deterministic approach based on dynamic programming called DTW and (b) the other is a statistical approach where the speech is modeled as a hidden Markov model (HMM). In our implementation we use both these to compare and hence verify the validity of the speaker. The use of multiple parameters (LPC, MFCC); multiple models (DTW, HMM; deterministic, statistical) give the speaker verification system robustness for verifying the identity of a person.

DRIVER FATIGUE DETECTION

Driver fatigue is one of the major causes of road accidents particularly on long highway drives. In general, there are six main causes of driver drowsiness, namely, poor sleep, sleep disorder, stress, monotonous driving, work shift, time of the day. Drowsiness detection techniques can be classified into three major categories, namely, sensor technology, computer vision technology and monitoring vehicle behavior. In [8] a method has been proposed to detect driver fatigue by measuring the duration for which the vehicle is at rest. The steering input behavior of the driver is monitored during a specified period of time [9] to detect the driver drowsiness. The lane tracking ability decreases as the time on task (Monotonous activity) increases [10]. Since time on task is directly related to sleepiness, there exists a correlation between lane tracking and sleepiness. The vehicle lane position could be used to detect drowsiness [11]. Normally when a driver sleeps, he tilts his head. In [12], a head tilt detector sounds an alarm when the head nods too much. In [13], EOG and Head Nodding (HN) signals are used to detect drowsiness. Many researchers dealing with yawn detection have focused their methods on geometric features of the mouth [14]. In [15], drowsiness is detected by mainly finding the characteristics of the eyes. The breath depth [16] is also used to detect the drowsiness. We propose a driver fatigue detection system

based on driver drowsiness using an infrared camera. The localization of face followed by the characteristics of eyelid movement is used classify the duration of the sleeping and normal driver. An alarm system based on the classification enhances the road safety.

The system uses an IR CCD camera positioned on the dashboard in front of driver's face. The IR illumination creates the bright pupil effect that creates a nearly perfect circle. The system searches for the eyes in each frame. If the opened eyes are not found for a few consecutive frames, the system draws the conclusion that the driver is drowsy and issues a warning signal.

If V be the IR video, then to proceed with the binarization we employed mean removal techniques as in Equation 1.

$$R[i, j] = V[i, j] - \bar{V} \qquad (1)$$

From this resultant $R[i, j]$ if we plot the histogram and take the peak, we will find the binary face with pupils as shown in image (1) of Figure 4. It also reveals some non-facial region along with face. To get the exact face we use the Euler number as in Equation 2 to get the component which has two or more holes in it which is further verified using template matching.

$$E = C - H \qquad (2)$$

where E is Euler number, C is the number of connected components and H is the number of holes in a region. Image (2) in Figure 4 shows a typical Euler image after applying Equation 2.

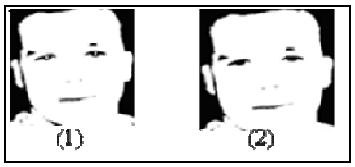

Figure 4. Different pupil states

Pupils can be found out by XOR-ing images (1) and (2) of Figure 4. Once the pupils are found they are tracked using nearest neighborhood techniques around them. An alarm will be raised if there are no pupil found during tracking for some frames, which directly signals that the eyes are closed or about to close, i.e. the driver is drowsy.

CONCLUSION

In this paper, we have described a few techniques for multimodal interactions with in-vehicle equipment for enhancing driving comfort, security and safety. While we have achieved a good performance of the algorithms in lab environment, a field testing and usability experiments are pending.

REFERENCES

1. VACC Submission, Driver Distraction prepared for the parliament of Victoria Road safety committee. (2005)
2. Lansdown, T. C., Brook-Carter, N. & Kersloot, T. Distraction from Multiple In-Vehicle Secondary Tasks: Vehicle Performance and Mental Workload Implications in Ergonomics, Vol. 47, No. 1, 91–104 (2004).
3. NHTSA, Drowsy driver detection and warning system for commercial vehicle drivers, Field proportional test design, analysis, and progress, *National Highway Traffic Safety Administration,* Washington DC. (2007)
4. Weirwille, W. W. Overview of Research on Driver Drowsiness Definition and Driver Drowsiness Detection, *14th International Technical Conference on Enhanced Safety of Vehicles,* 23-26 (1994).
5. Auto Theft Statistics URL: http://www.auto-theft.info/Statistics.htm, last retrieved on 26th Nov 2008
6. Malima, A., Ozgur, E., & Cetin, M. A Fast Algorithm for Vision Based Hand Gesture Recognition for Robot Control, *14th IEEE Signal Processing and Communications Applications*. (2006)
7. LaViola, J. J. A survey of hand posture and gesture recognition techniques and technologies. Technical Report CS-99-11, Department of Computer Science, Brown University. (1999)
8. Seko, Y., Iizuka, H., Yanagishima, T. & Obara, H. Method and system for detecting driver fatigue including differentiation of effects of rest periods, US Patent 4602247 (1984)
9. Ferrone, C. W. & Sinkovits, C. System and Method for Monitoring Driver Fatigue, US Patent 7427924. (2005)
10. Dureman, E., & Boden, C. Fatigue in Simulated Car Driving, *Central Research Institute*, Nissan Motor Company, 547-554. (1972)
11. Skipper, J., H., Wierwille, W., & Hardee, L. An Investigation of Low Level Stimulus Induced Measures of Driver Drowsiness. *Virginia Polytechnic Institute and State University IEOR Department Report* 8402, Blacksburg, VA. (1984)
12. Kyrtsos & Christos, T. (Southfield, MI), Drowsy driver detection system, assigned to Meritor Heavy Vehicle Systems, US Patent 5900819. (1998)
13. Hussain, A, Bais, B, Samad, S. A. & Hendi H. F, Novel Data Fusion Approach for Drowsiness Detection, *Information Technology Journal*, Vol 7, 48-55. (2008)
14. Fan, X., Yin, B. & Sun, Y., Yawning Detection for Monitoring Driver Fatigue. *IEEE Machine Learning and Cybernetics Intl. Conf.* (2007)
15. Zhang Z., & Zhang, J., Driver Fatigue Detection Based Intelligent Vehicle Control. *18th Intl. Conf. ICPR* (2006)
16. Ikegami, T., Nanba, S., & Yanai, K., Drowsiness detecting apparatus and method, US Patent 7397382 (2008)