Towards Multimodal Human-Machine Interface for Hands-free Control: A survey

Technical Report: CES–510

Lai WEI and Huosheng HU

School of Computer Science & Electronic Engineering
University of Essex, Colchester CO4 3SQ, United Kingdom

January 2011

Table of Contents

Introduction

1.1Hands-free control

For decades, a key issue that lies in science and engineering has been how human could interact with a computer or machine in a natural way. Until now, the most successful and widely used interfaces are keyboard and mouse through which users can operate a computer or control various devices that are connected to a computer. As human hands have a complex anatomy structure which could make delicate gestures and precise control movements, hand movements are one of the most common ways that are applied in HMI applications. However, for people with severe physical disabilities such as spinal cord injury, quadriplegia, hemiplegia or amputation, keyboard and mouse are no longer adequate. It is necessary to develop novel human-machine interfaces for disabled and elderly people to use computers and robots for the better quality of life in the society.

Hands-free control is an important concept in designing human machine interfaces for people with disabilities. It focuses on setting up communication between a machine and a part of human body such as face, shoulder, and limb via a series of kinetic movements. It can replace normal hand movements. Apart from kinetic movements, there are many other forms of human body movements that can be deployed for building novel HMIs, such as facial expression, eye movement, voice, body and limb movements etc. Fig. 1 shows that a novel HMI based wheelchair has been created for Dr Hawking to gain necessary mobility, who is severely disabled by motor neuron disease (amyotrophic lateral sclerosis or ALS).

In the past decades, human machine interfaces have been under extensive development by using electro biological signals such as brain signals (EEG), muscle signals (EMG), and eye retina signals (EOG) and by using human head gesture, facial expression, or limb gestures etc. A new trend for resolving human intension lies in computer vision which uses an image based method to analyze human body movements, head gesture movements or facial expressions.

Fig. 1 Dr. Stephen Hawking sitting on a wheelchair

It is estimated that nearly a half of the people over 80 years old need extra help from a rehabilitation tool, and 40% of the people with disabilities are resulted from disabled hand movements. It is necessary for developers and researchers to investigate substitute means for people with different level of disabilities and also for those who has special needs according to category of diseases [11] such as spinal cord injure, Parkinson's disease, muscular dystrophy, arthritis, quadriplegic, amputation etc. Whereas the symptoms of a full list of disabilities are describe in [11] in details, personal damage and disability situation caused by the same disability genre may vary individually from person to person. And a special interface is needed to customize for each person based on their individual situation, the complexity of building such hands-free systems can vary according to the level of disability and individual situation.

Today, hands-free control is widely used to help disabled people to control a rehabilitation device or a wheelchair. And the prospective of application is expanding into more and more areas. Hands-free control functions are not only helpful to elderly and disabled, but also useful for normal people who may need to use their hands for other tasks at the same time. Therefore, normal people are capable of manipulating multiple tasks at same time.

1.2 The way to implement hands-free control

The engineering endeavor to provide communication abilities to aged people and people with disabilities have resulted into many ways to realize hands-free control by employing various movement of the human body. Simulating a switch function is most simple and popular ways to implement hands-free control. Control movements are originated from tongue, lips, neck, speech, breath or facial expressions, muscle contractions (shoulder muscle, limb muscle or foot muscles etc.), and can be applied to activate a simulated switching movement.

One of the most popular commercialized switching based hands-free control devices is sip and puff interface which is shown in Fig. 2. The interface can simulate the function of turning on and off a switch by sipping or puffing the air into a mouth tube, and a switch action is activated by the change of the air pressure in the tube. Mouth joystick is another simple and low cost mouth mounted hands-free control interface which allow the user to control a joystick by griping the stick with mouth. These interfaces are the earliest forms of hands-free control interfaces which enable disabled people to control devices such as a powered wheelchair or a computer.

Since both contact switch button and sip & puff interface requires a user to have physical contact with the control interface. As a replacement of contact movement based interfaces, researchers looked for non-contact interfaces to track human movements to implement hands-free control using head movements, facial expressions, eye movements. Head gesture based HMIs use some form of head tracking technique as a vital part of design are of particular benefit to the disabled. People who have considerable difficulty in using conventional input devices can use head gestures for hands-free control purpose. Head gestures are normally tracked by a camera with an image processing technique. There are two basic ways for tracking head movements with a camera: one is using artificial label which is deliberatively attached to user's head or face area, and another is synthesizing geometric or color features of head or face for tracking the correspondent head or facial movement. Fig. 3 (a) shows a hands-free mouse SmartNav [48] Sip and Fig. 3(b) shows the puff interface used for controlling an electrical powered wheelchair.

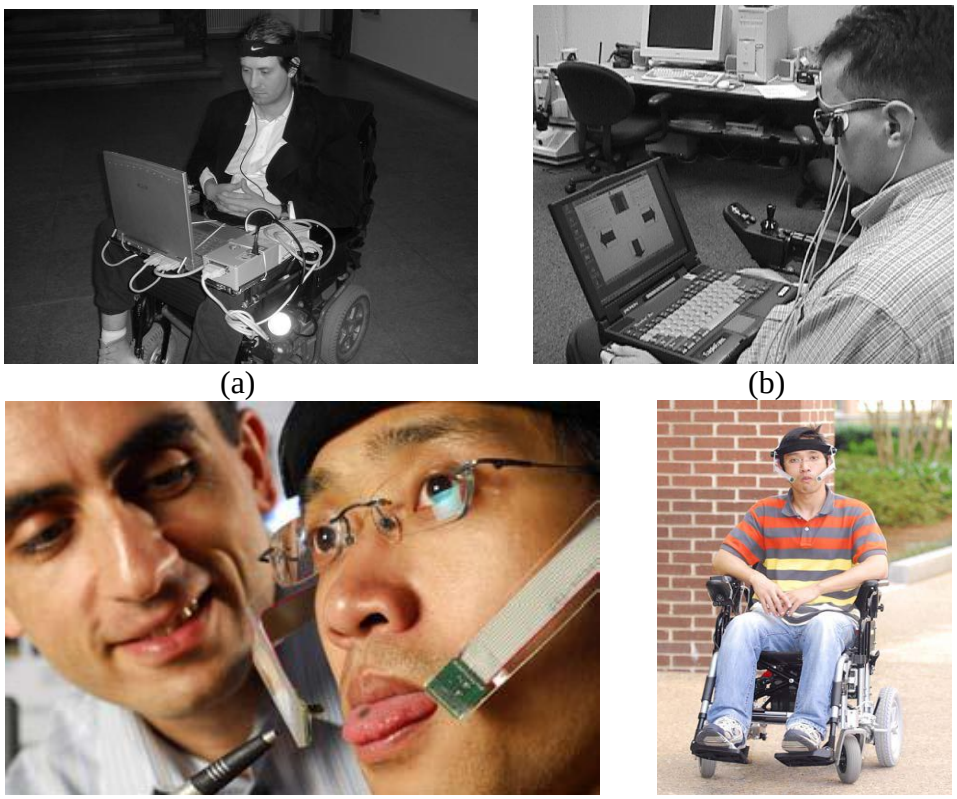
(a) (b)

Fig. 2 (a) Sip and puff pipes and switch sensor box; (b) Sip and puff interface for controlling an electrical powered wheelchair.

(a) (b)

Fig. 3 (a) SmartNav [48] interface for hands-free cursor control of a personal computer; (b) Vision based interface for controlling an electric wheelchair.

Head gesture HCIs use some form of head tracking as a vital element of their design. Such interfaces are of particular benefit to the disabled, who have considerable difficulty in using conventional input devices. Vision based head gesture interface recognition, as a way for hands-free control of powered wheelchair, is under extensive research [27] [50] [52] in the past decade. Different methodologies of tracking head or face features are proposed and the control strategies are tested on several "intelligent wheelchair" platforms. Fig. 4 (a) shows the forehead EMG based control interface [7] [14]; Fig. 4 (b) shows that EOG signals obtained by circumocular electrodes are used to control a wheelchair [8]. Fig. (c) and Fig. (d) show a tongue movement based HMI [12] for controlling a powered wheelchair.



(a) (b)

(b) (d)

Fig.4 Hands-free control of powered wheelchairs.

Two non-invasive BCI systems using Electroencephalography (EEG) signals are shown in Fig. 5, which is a new subject for human computer interaction study in recent years. The users may imagine some movements used for controlling a computer or a robot, which will stimulate a neurological phenomenon and change the electricity distribution around the scalp. EEG make use these evoked electricity potentials distribution activity and the surface electricity pattern during imagination process are collected by a computer as the clue directly connected to user's imaginary control movement.

(a) (b)

Fig. 5 Brain computer interface (BCI) (a) BCI caps with flexible electrodes used to attain EEG signal applied in research experiments; (b) Commercialized BCI headset with fixed electrodes and its screen interface [15].

So far, experimental BCIs have been used for controlling a cursor on a computer screen or moving a graphical object through a virtual environment, although using BCI as a means to control rehabilitation devices such as a wheelchair has been experimented and propose in [14] [49], and a couple of efforts and improvements have been made in terms of better EEG acquisition techniques and more advanced classification algorithms. It is arguable that the current scalp potential based BCI interfaces are suitable for controlling a real time control system such as a wheelchair device due to the current characteristics of the EEG control system such as slow responses and low recognition rates, and EEG pattern induction and stimulation process often need a user's full concentration of attention.

Emotiv [15] is among the first commercialized EEG based BCI devices used for game and research purpose. It has a headset with fixed number of EEG channels and allocation of electrodes for different users. As shown in Fig. 5(b), the product contains its own EEG interface to train and customized imaginary BCI interface for its user in which the user can rotate or remove a 3D object by brain imaginary activities.

Hands-free tongue interface for disabled people also reported in [12], in which a magneto-inductive sensor is attached on the surface of the tongue to track the tongue movement in the mouth. Huo et al. used the tongue interface to control a wheelchair, the experiment result shows that the proposed tongue drive system is helpful to people with severe disabilities to control their environments, access computers, and drive powered wheelchairs.

It is clear that a number of hands-free control HMI systems have been developed in the last decade due to the advancement of electronic, computing and sensory systems. Although these systems can work well for a certain group of disabled people, they may not work at all for the people with different disability. Single modality HMIs, either visual-based or biosignal-based, may not work well since each of them observes only partial information of face movements. Therefore, it is necessary to study how to effectively combine different modalities to form a multimodal system in order to cope with difficulty that a single modality cannot handle.

1.3 Introduction of multimodal approach

Multimodality is a ubiquitous phenomenon in nature. Animal interacts with each other using odour, crowing and posing; insects communicate with each other using pheromone, body movement, antenna contact, etc. Multimodal is a natural way of human expression too. In the real world, we express ourselves by using face expression, voice, eye contact, body language, limb gesture etc. It is clear that human interaction is a naturally multimodal process and contains deep psychological and physiological expressions; even the same expression can have different means under different circumstances. There is not a certain rule for human or artificial intelligence to formulate its modal.

The advancement of computing and pattern recognition techniques enables a machine or robot to perceive and analyse human intentions. Also the machine learning techniques such as reinforcement learning can help the machine to learn from human and understand human intention better and better. There are no shortcuts to solve these problems as human expression is evolved from nature by millions of years. The only solution is to escalate the level of machine intelligence towards a higher lever of understanding, just like human does.

There are a wide range of multimodal or multisensory systems that are applied in a variety of areas from security surveillance to environment monitoring and from supply-chain management to health care and rehabilitation. Therefore, the multimodal approach can be applied to many real-world applications for the benefit of the better performance and the high reliability, including face detection and tracking, face recognition, pedestrian or human detection, surveillance, speaker detection, user authentication, interactive and smart environments, human robot interaction [47] [52].

2 Formulation of a Multi-modal HMI

2.1 Architecture of a multi-modal HMI

They are a new class of emerging systems that aim to recognize naturally occurring forms of human language and behavior, with the incorporation of one or more recognition-based technologies (e.g., speech, pen, vision). Multimodal interfaces represent a paradigm shift away from conventional graphical user interfaces. They are being developed largely because they offer a relatively expressive, transparent, efficient, robust, and highly mobile form of human-computer interaction. They represent users' preferred interaction style, and support users' ability to flexibly combine modalities or to switch from one input mode to another that may be better suited to a particular task or setting.

Human body is a multi model system. In our daily life, we express our intension by various means. For example, we begin to cry when we born, we learn to make facial expressions and gestures when in the cuddle, and gradually we learned talking, more complicated facial expression and complications, drawing and walking afterwards. All these ways of learning and communicating contribute a crucial way and make us a complete human being. Therefore, human beings are a complex system, i.e. a multi-modal intelligent system. The communications between machines and humans have been a hot research issue for decades. Researchers are currently seeking a more natural and pleasant way to communicate with machines or robots.

Another issue is how a machine or a robot can understand human intension throughout multiple human expressions. When a person is talking, he or she not only uses language, but also other assistive ways such as facial expressions, hand gesture, head gesture, body language etc., as shown in Fig. 6. So it is necessary for a machine to understand human intension throughout multiple activities, which covers a wider range of understanding by using multiple sensors or fusion of different methodologies such as EMG, EOG, EEG, computer vision etc.

Multimodal perception has been deployed in mobile robots perception and navigation for few decades. Typically, a mobile robot can collect information from the surrounding environment by using a series of sensors, such as sonar, laser range finder, camera, microphone, odor sensor, etc. Recently, multimodal perception has been used in human-machine interfaces since service robots need to understand human intentions in order to provide necessary service in time. Therefore, human is an important factor that robot needs to learn and deal with. Service robots should be a friend of users and has an effective way to communicate with users; otherwise no people would like to use the robots.

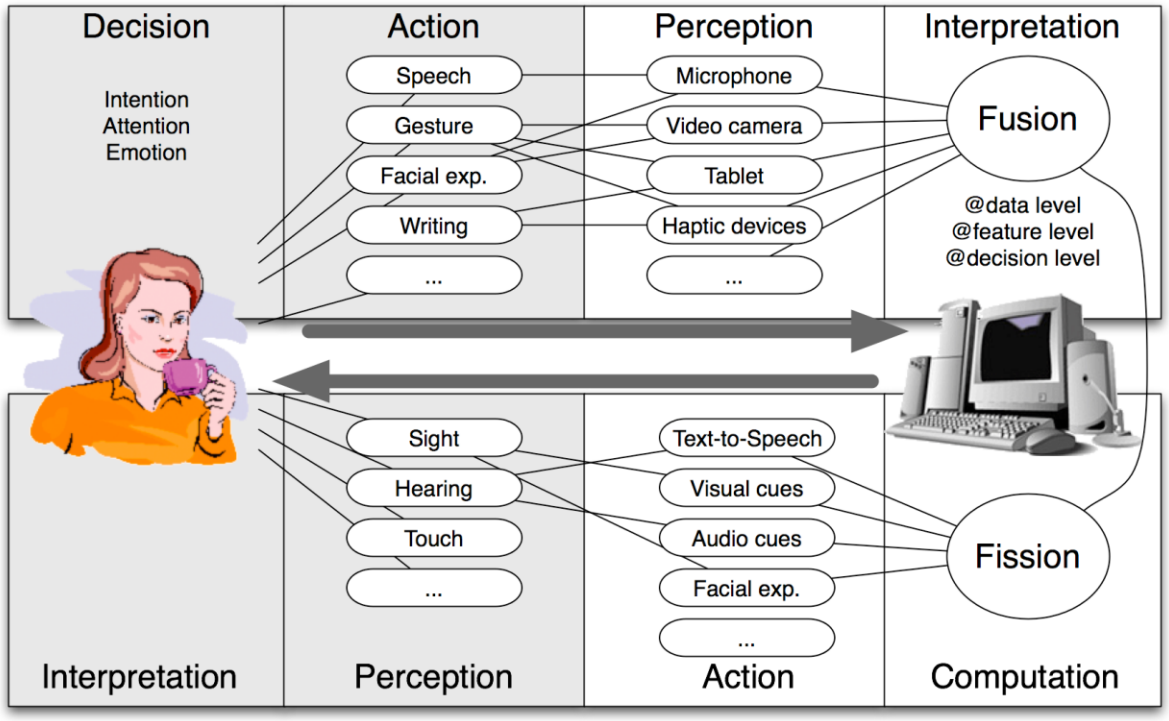
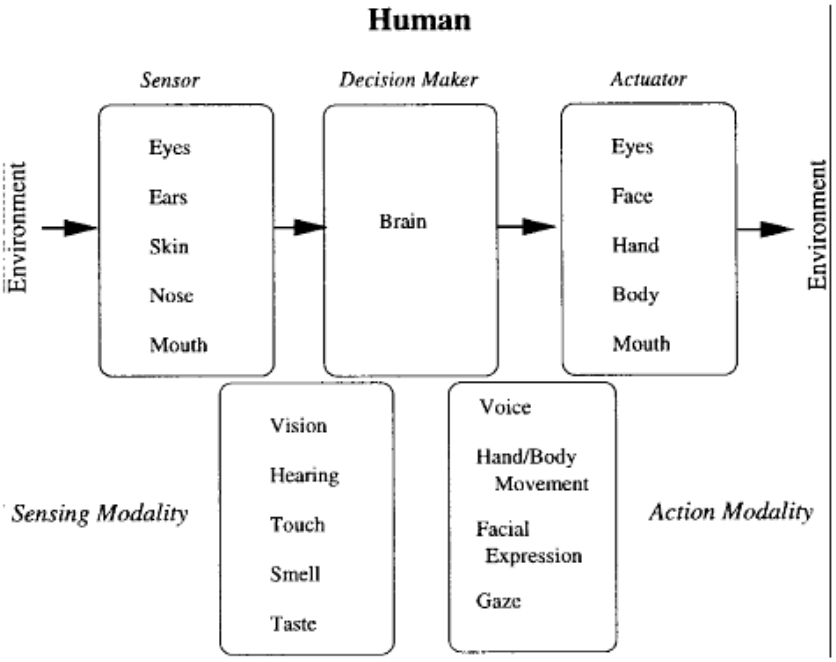


Fig. 6 Architecture of multimodal man machine interaction [17]

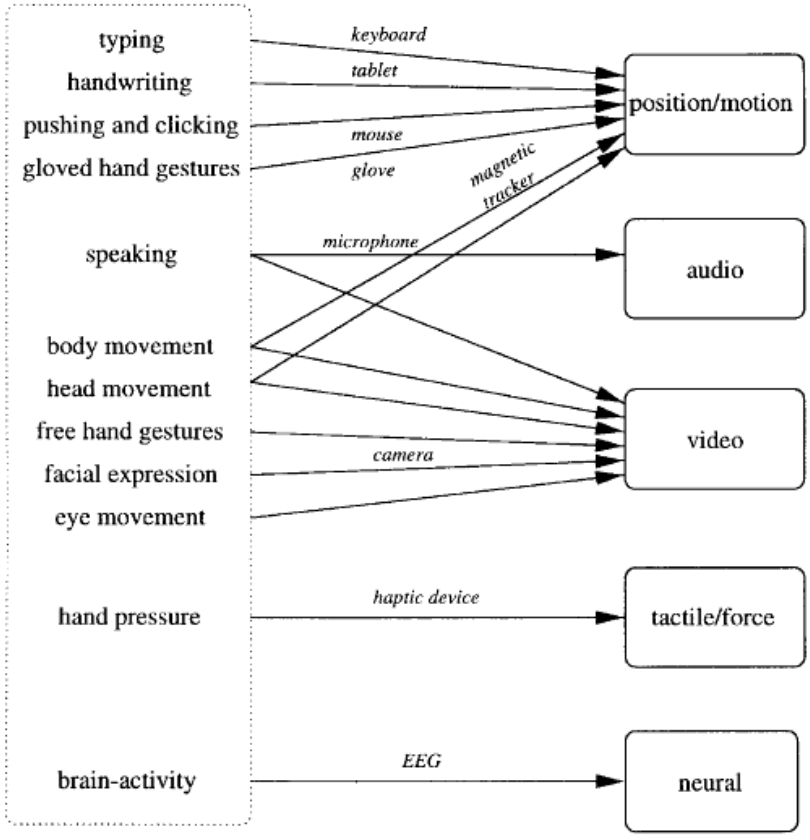
The most common multimodal interfaces combine different input modes such as speech, pen, touch (tactile), gestures, eye gaze, head and body movements. The advantage of using these multi inputs is that the multiple modalities increased usability, and provide the user with more possibilities to maximize their communication abilities. The use of multimodal perception can have two benefits. One is to increase the recognition accuracy by employing more information, such as speech recognition by lip reading using combined vision and voice recognition systems. Another advantage is to enrich the movement patterns and therefore make different combination of joint movements into one integrated system.

2.2 The way to implement multimodal HMIs

Multimodal HMI systems have evolved rapidly during the past decade, with steady progress toward building more general and robust systems. Major developments have occurred in the hardware and software needed to support key component technologies incorporated in multimodal systems, in techniques for fusing parallel input sensory data, in natural language processing (e.g., unification-based integration), and in time-sensitive and hybrid architectures. To date, most current multimodal systems are bimodal, with the two most mature types involving speech and pen or touch input, and audio-visual input (e.g. speech and lip movements). However, these systems have been diversified to include new modality combinations such as speech and manual gesturing, and gaze tracking and manual input.



(a)



(b)

Fig. 7 (a) Human cognitive system; (b) The way to sense human perception [2]

Multimodal HMI applications include map-based and virtual reality based systems for simulation and training, multi-biometric person identification/verification systems for security purposes, and medical, educational, and web-based transaction systems. Given the complex nature of users' multimodal interaction, cognitive science has played an essential role in guiding the design of robust multimodal HMI systems. The development of well integrated multimodal HMI systems that blend input modes synergistically depends heavily on accurate knowledge of the properties of different input modes and the information content they carry, how multimodal input is integrated and synchronized, how users' multimodal language is organized, what individual differences exist in users' multimodal communication patterns.

In [17], a general architecture for multimodal HMI is proposed in a cycle structure. Dumas et al. divide the human machine interaction procedure into four different states, namely decision stage, action stage, perception stage and interpretation stage. In the decision state, the communication message content is prepared consciously for an intention, or unconsciously for attention content or emotions. In the second state, the communication means to transmit the message being selected, such as speech, gestures or facial expressions. The machine, in turn, will make use of a number of different modules to grasp the most information possible from a user.

In [2], a structural modality of human sensing and action is proposed. As shown in Fig. 7(a), human cognitive and perception system are divided into three layer hierarchical structures, namely sensor level, decision level and actuator level. The first layer is human senses which can perceive environmental information by vision, hearing, touch, smell and taste. The decision level is the human brain and human consciousness which is a buried and hidden phenomenon. In the third actuator level, the human intension is interpreted by action movements such as voice, hand/body movements, facial expression and gaze. In Fig. 7(b), Jaimes, et al. further connected a complete sensible range of human actions with corresponding machine based sensing and inspection methodologies.

Jaimes and Sebe proposed these three levels for fusion for building multimodal interface [2]. As depicted in Fig. 8, multimodal data can be processed at three levels which are data level, feature level and decision level. Data-level fusion is normally used when dealing with multiple signals coming from a similar modality source such as using two webcams recording the same scene from different viewpoints. With this fusion scheme, no loss of information occurs, as the signal is directly processed. Due to the absence of pre-processing, it is highly susceptible to noise and failure.

Feature-level fusion is a common type of fusion used for tightly-coupled or time synchronized modalities. The standard example is the fusion of speech and lip movements. Feature-level fusion is susceptible to low-level information loss, although it handles noise better. The most classic architectures used for this type of fusion are adaptive systems like artificial neural networks, Gaussian mixture models, or hidden Markov models. The use of these types of adaptive architecture also means that feature-level fusion systems need numerous data training sets before they can achieve satisfactory performance.

Decision-level fusion is the most common type of fusion in multimodal HMI applications. The main reason is its ability to manage loosely-coupled modalities like, for example, pen and speech

interaction. Failure and noise sensitivity is low with decision-level feature, since the data has been preprocessed. On one hand, this means that decision-level fusion has to rely on the quality of previous processing. On the other hand, unification-based decision-level fusion has the major benefit of improving reliability and accuracy of semantic interpretation, by combining partial semantic information coming from each input mode which can yield "mutual disambiguation". Fig. 9 shows a multimodal interaction environment which are used to describe the future prospective of multimodal human to computer and human to human interaction [16].

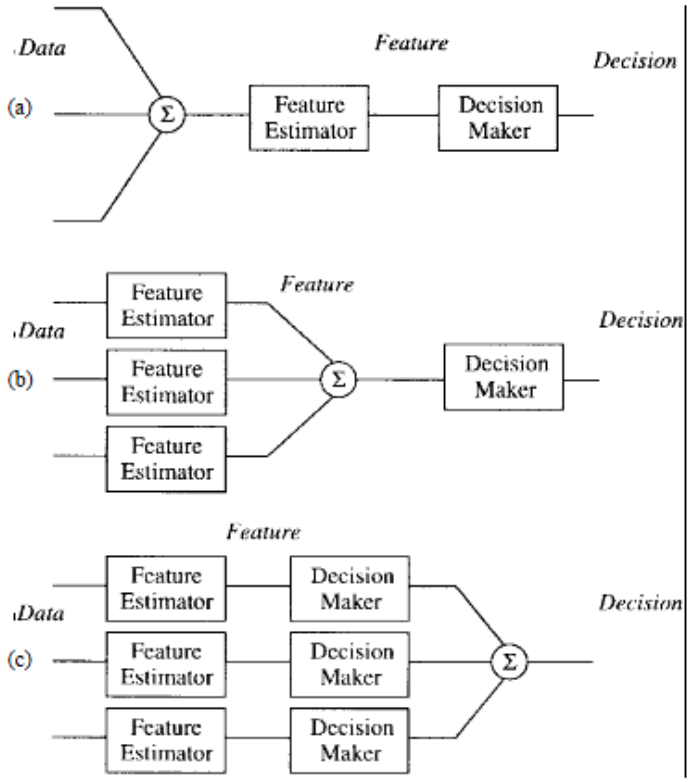
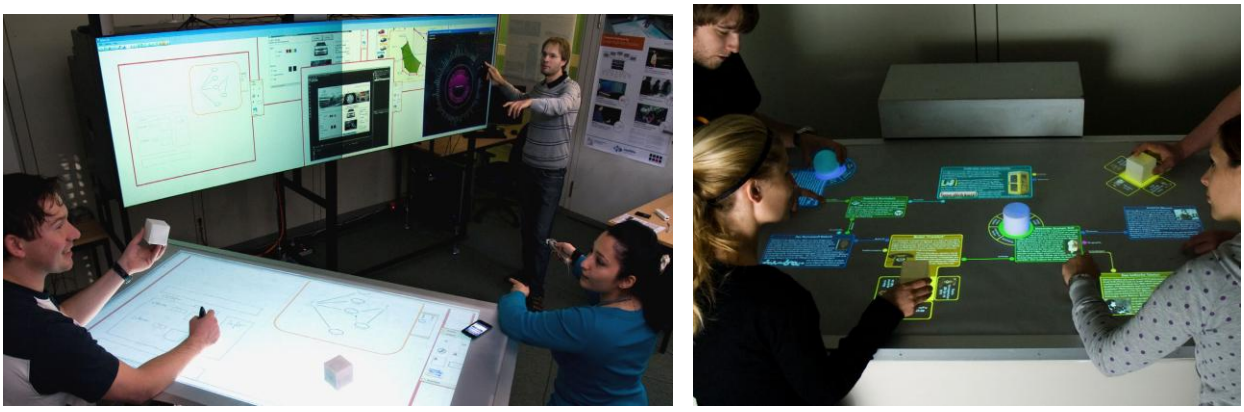


Fig. 8 Diagram showing the fusion of multimodal system in three layers namely data fusion layer, feature fusion and decision fusion [2].



1) (b)

Fig. 9 (a): A smart environment for multimodal interaction in an intelligent conference; (b) Multimodal HCI for a human-human and human-computer interaction in a conference application. [16]

3 Implementation of Multimodal HMI Systems

The development of computer and manufacturing in the past years brings opportunities for the development of multi sensory systems with large amount of data processing requirement. As a new applicable area that begins to thrive in the near future, most of the work involving multi modal interaction stayed on the concept layer, and in the past few years, many advanced multimodal HMIs have been developed especially in the virtual reality area.

The state-of-the-art research progress in the MMHUI area can be divided into two directions. One is the human centred direction, which sees human body as a multimodal system and explores HMI based on applicable human movements including facial expressions, human affections, head gesture, body gesture, eye gaze, muscle movements, brain activities and so on. The other is the application based direction, which concentrates on the applications that have been developed using multimodal interfaces, including web browsing, virtual reality interaction, human robot interaction, rehabilitation, assisted devices for disabled and elderly etc. By doing this, we can cover a wide range of prospective and view closer comprehensive to the development of MMHCI. However, multisensory and multimodal systems are a big family that has wide applications.

The recent research on multimodal HMIs has been surveyed in several articles. Since multimodal HMI interfaces are still under the earlier development, successful HMI interfaces are very few. In this report, our main focus is on the computer vision based systems that used for analyzing human expression and human emotion, and biometrical signal based systems that used for human inspection etc. In this way, we can envisage a wider range and grasp the tendency of the development in this area that can extend into area into developing new human machine interfaces.

3.1 Facial expression and human emotion analysis

Facial expression and human emotion has been involved into many HMI applications. In [34] [19], facial expressions are used to analyze and detect potential dangerous of the drowsiness of driver. Automatically detected human affections are also used to analyze underlying autism tendency of children and polygraph are been used for detecting the potential crime intentions. In the robotics area, human expressions are recognized and simulated by avatar, and these applications have prosperous applicability into human computer interactions and enhance the human to human communication as well. In a multimodal recognition system, human expression clues are normally used as a factor that can be taken into account to deduce and judge human emotion states.

In [33], Ekman and his colleagues concluded that six basic emotions that can be universally recognized: happiness, sadness, surprise, fear, anger and disgust. These six “universally distinguishable” emotions have been studied and formed the foundation theory for machine or computing based analysis and recognition of human emotion and affection. As an underpinning index of the human expression, human affection is closely connected with human facial expression clues and body gesture cues. Besides, human affection can be analyzed by a series of substantial physiological and bio potential signals such as brain signals measured via functional Near Infrared Spectroscopy (fNIRS), scalp signals measured via electroencephalogram (EEG),

and peripheral signals, namely, cardiovascular activity, including Electrocardiography (ECG), blood pressure signal; electrodermal activity such as skin conductance or galvanic skin response (GSR), electromyogram (EMG) activity (from corrugator supercilii, zygomaticus, and upper trapezius muscles).

Currently, human facial expression is viewed as multimodal phenomenons that are connected with facial experience, and bodily gestural expression, audio or acoustic utterance etc. Human affection is the inner feeling of people and can be detected by synthesizing a combination of biometric phenomenon and facial expression. Human affection can be jointly judged by body gestures too. In computing intelligence based recognition for both facial expressions and human affections, different methodologies are being applied and human expression is acted as an assistive and complementary tool for analyzing human emotion.

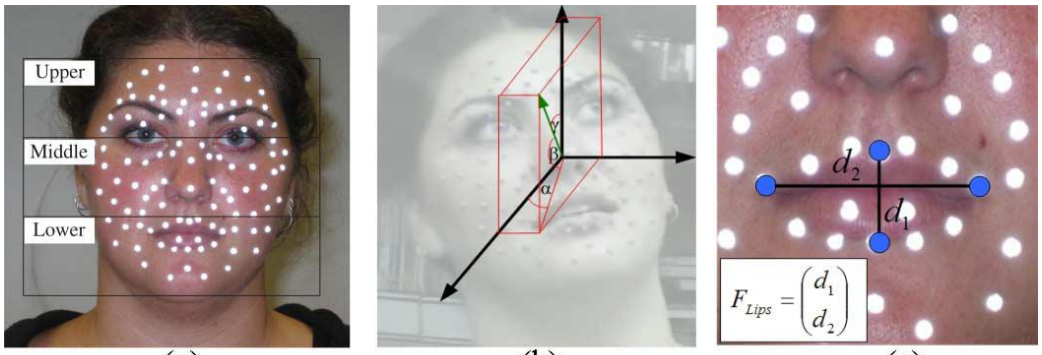


Fig.10 Facial gestures are tracked by markers allocated above. And in the experiment in [3], the whole face area is divided into three layers.

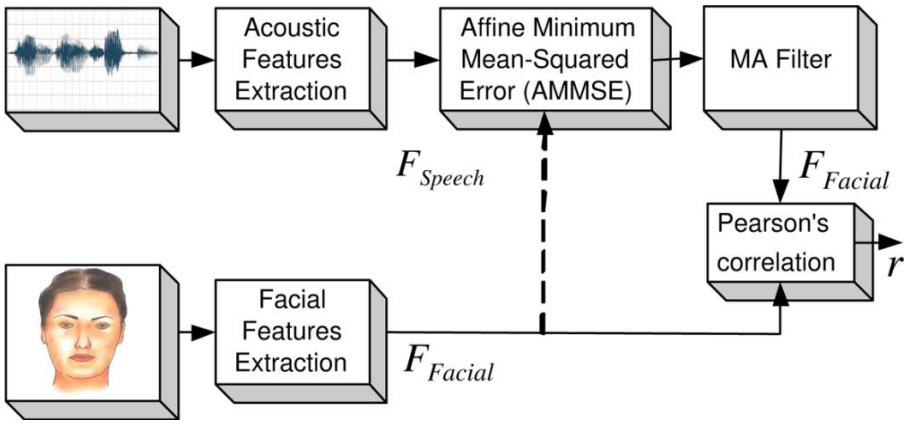


Fig.11 Schematic graph in [3] showing the process of synthesization of audio-visual information and analyzing level of correlation under four designated emotional conditions.

Human voice and facial expressions are internally and intricately connected by human emotions. The correlation between facial gestures and speech are normally investigated and analyzed synchronously by synthesizing audio–visual database recorded from a subject who was asked to read sentences in different emotion conditions. Busso and Narayanan [3] analyze the correlation

between facial gestures and speech using four emotional states: neutral, sadness, happiness, and anger. The facial gestures are collected using parameterized face markers as shown in Fig. 10. A multilinear regression framework is used to estimate facial features from acoustic speech parameters. The levels of coupling between the communication channels are quantified by using Pearson's correlation between the recorded and estimated facial features as shown in Fig. 11. The results show that the speech under the four emotional circumstances is strongly correlated with the appearance of the face.

The idea of getting face appearance by marking whole face area with detectable labels has a great deal in common with resolving facial geometrical features using computer vision. Computer vision, i.e. image information, is an important way to extract human facial expression and deduce human emotion. Instead of using two modals to analyze human emotion, Loic and Ginevra et al. recognize human emotion using a combination of three expression modalities which are facial expression, gesture and acoustic analysis of speech [4]. In their study, eight emotion states such as anger, despair, interest, pleasure, sadness, irritation, joy and pride are sampled from subjects from five different nationalities (French, German, Greek, Israeli and Italian). As shown in Fig. 12, the fusion of the three models is performed at the feature stage.

The feature data of the face applied in the experiment are the tracked points extracted from eyes, mouth, eyebrows and nose, which were detected by Viola Jones algorithm, i.e. a cascade of boosted classifiers trained with Haar-like features. The body feature data are five expressive motion cues which are synthesized from EyesWeb platform (Fig. 13) by tracking the silhouette of body and hands of subject. A set of 377 voice features based on voice intensity, pitch, Mel Frequency Cepstral Coefficient, Bark spectral are formed into voice feature data. Loic and Ginevra compared Bayesian classification result among three single modalities as well as double modalities matches namely "face-gesture", "face-speech" and 'gesture-speech' modalities. The result shows that three modal data have better classification accuracy compared single double modalities.

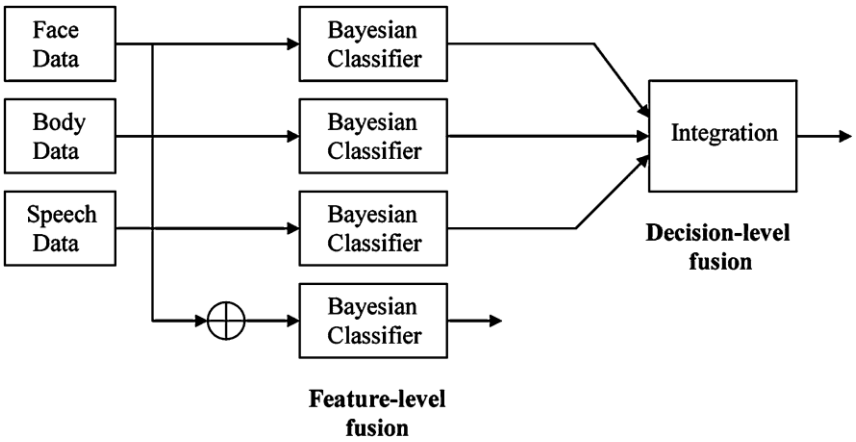


Fig.12 Fusion steps of three modalities which are face data body data and speech data represented in [4].

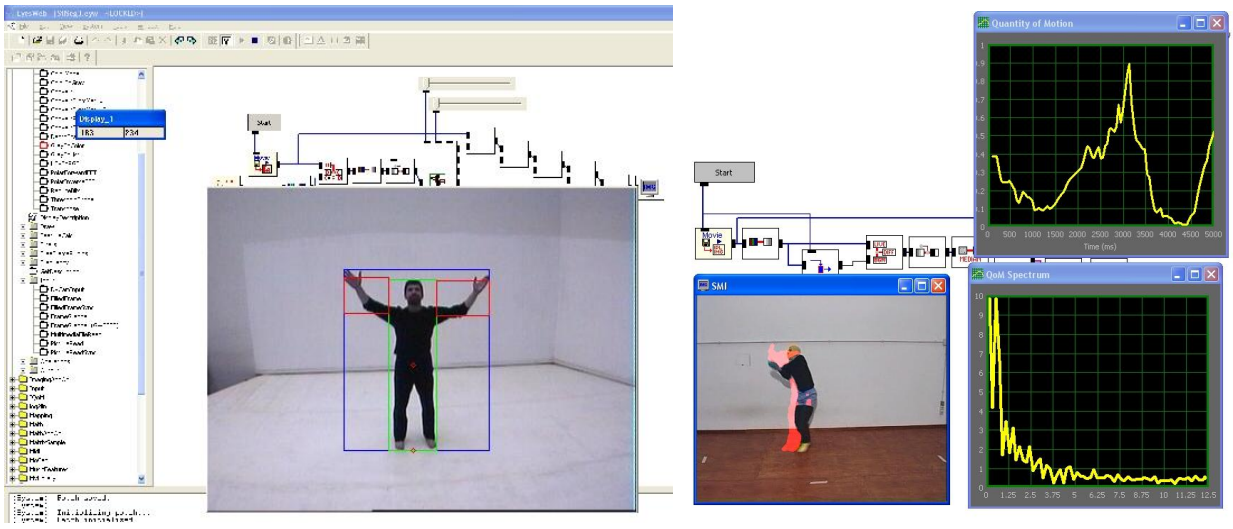


Fig.13 The EyesWeb interface [5] used in [4] to analysis human body motion expressions

Zeng et al. performed affect recognition on four cognitive affects: interest, boredom, frustration and puzzlement and seven basic affective states: neural, happiness, sadness, anger, disgust, fear and surprise of 20 random subjects [36]. A feature selection approach is used for selecting best relevant features set from facial expression images and acoustic speeches to select best classification accuracy from a Multi-stream Fused Hidden Markov Model (MFHMM) approach. The experimental results show that bimodal affect recognition can gain 7.5% and 6.1% classification accuracy bonus in person-dependent test and in person-independent test respectively compared with the best result achieved with the signal modal approach. And the experiment also envisaged that MFHMM can outperform traditional HMM in fusing multi-stream affect information.

In [41], Gunes and Piccardi combined facial expression and affective upper body gesture information tracked from a camera at both feature and decision levels. They construct a feature set which composed of 152 face features and 170 upper-body features for implementing an automatic visual recognition system. The classification results show that the system with two modalities (facial expression and upper body gesture information) achieves better recognition accuracy in general, compared with classification using one modality alone;

Human voice, facial expressions or gestures can be monitored from outside. However, bio-signals within human body can be inspected for detecting physiological phenomenon of the human emotion. Synthetic biometric data are widely used as a way to predict human intensions. These signals such as EMG signal from designated muscle, and EOG signal from eye movement's states, as well as EEG signals from activities of the human brain can contribute to estimating the emotional state and therefore deduce the human intension. Also, these biometric signals can assist the recognition of human emotion state with the other existing modalities such as facial expression, gestures and voice.

Takahashi proposes an emotion recognition system based on multi-modal bio-potential signals [45]. EEG signal, blood pressure signal, and skin conductance are collected from 12 subjects.

Commercial film sections from TV are used to stimulus five emotions: joy, anger, sadness, fear, and relax. Six statistical features including means, absolute value and standard deviations are extracted from raw sensor data and support vector machines (SVMs) are applied to classify five emotions. Savran et al. [46] uses three modalities: (i) brain signals obtained from fNIRS methods which is a spectroscopic neuro-imaging method for measuring the level of neuronal activity from brain; (ii) facial video and (iii) scalp EEG signals to detect three emotion states of five subjects. Fusion between fNIRS with facial video and EEG with fNIRS are carried out seperately.

With a valence and arousal emotion modal, subjects are asked to quantitative their emotions with score from one to five. Active appearance models (AAM) were used for extracting facial features based on an active contour-based technique and a Transferable Belief Model (TBM) was applied for data classification. Emotional stimulus is given for acquiring emotion data. Kim et al. [32] uses bio-potencial signals from three modality signals which are electrocardiogram, skin temperature variation and electrodermal activity from the surface of human skin and used them to recognize two categories of human emotions each contain three and four of emotions respectively. The recognition result performed by SVM classifier using cross-validation technique over 50 subjects' shows that the recognition correctness rate of three and four of emotion groups can reach 78.43% and 61.76% f respectively.

From the above multimodal system paradigms, we can see that in terms of detection of human emotion:

1) Under the same experimental condition and recognition methodology, multimodal systems have better classification accuracy than single modal system or system that has less modalities;
2) Facial expression is a basic structure that contains most relevant information that indicates human emotion, and vision based facial expression tracking technique are an common and convenient way for getting facial expression information;
3) Support vector machine and hidden Markov modals are good classification tools and multi-stream data fusion analyzers that are competent for synthesizing and building multimodal systems; and
4) The features selected for representing each modal and the level in which multimodal systems are fused are two main factors that can affect the classification accuracy and system performance. For each system, a special recognition system structure is designed. Selecting best features and multimodal fusion strategies for designed multimodal systems are still an issue to be investigated.

Since human emotion and expression is very complicated and difficult to model, a robust multimodal HMI is required. Through the construction of multimodal HMI interfaces, a machine or robot can understand human intention better, and becomes more "intelligent".

3.2 Eye movement and gaze tracking

Pastoor et al. created a multimedia system including a multimodal visual operating 3-D display by eye-controlled interactions [25]. A computer vision based eye gaze tracker and a head tracker are employed into the interface. The user can interact with a 3D display by simply looking at the object and head tracker can recognize head movement to open the view of a document the user

gazed at. Based on the proposed system, more input modalities such as keyboard and hand gestures can be added to enrich the interaction performance; Fig. 14 shows a complete system prospective.

Hands-free control can be implemented by various human movements that are recognizable to a computer by different sensors. Eye gaze analysis is a method that is widely used for human computer interaction purpose. NaviGaze is software that can implement hands-free control by user's eye movement and head gesture [38]. Both eye gaze and eye close movements are used as a cursor based computer interface which mimics the function of a mouse, eye closing movements are used as left or right clicking of the mouse.

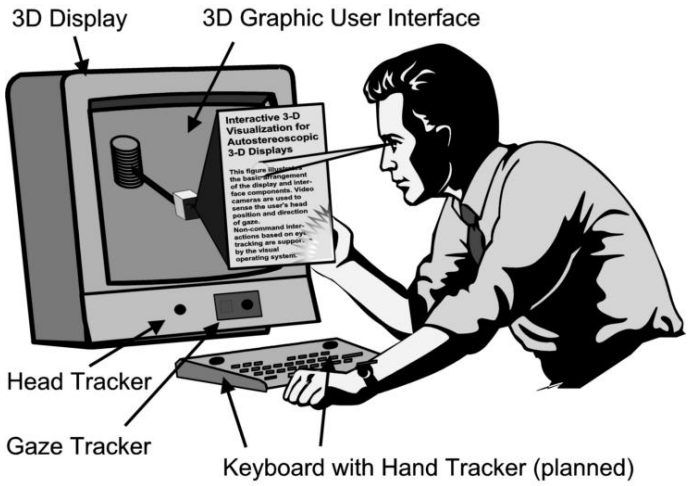


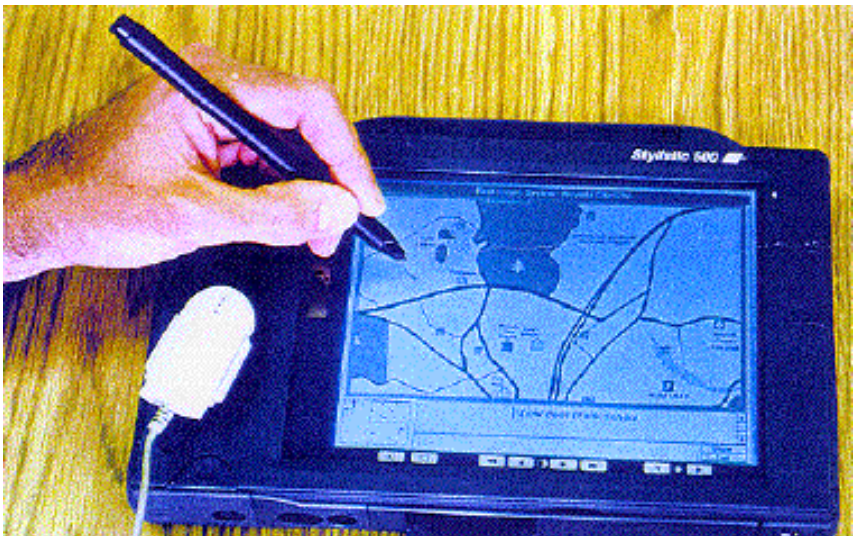
Fig. 14 Basic components of MMHCI system the proposed system [25]

3.3 Speech recognition and synthesis

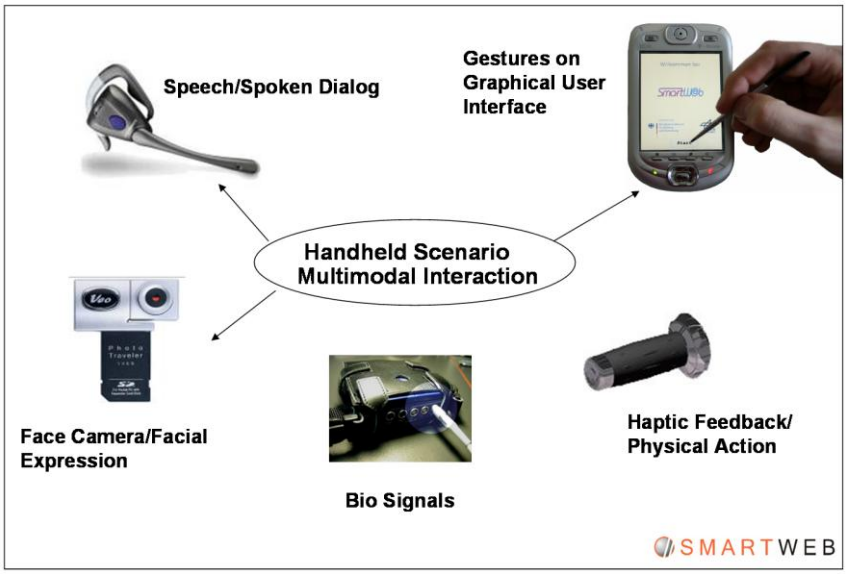
Bolt's "Put-That-There" system [10] is a pioneering speech input based MMHCI system that integrates arm gesture as an assistive mean of interaction. After that, a number of speech recognition based HCI interfaces such as speech and pen system [13], which incorporate a touch screen pen input modal are developed, and large number of successful applications of multi modal interface over single modality has been uncovered. Sonntag, et al. [30] proposed a dialog based multimodal architecture based on speech and pen inputs, which incorporates spoken dialog recorded by a Bluetooth module, gestures tracked with PDA touch screen, facial expressions from camera images for making a dialoged based multimodal web interface. Based on IBM's speech-to-text ViaVoice speech recognition engine, Perzanowski and Schultz et al. [47] developed a voice command based multimodal human-robot interface incorporating gesture inputs. The interface enables a user to interact with mobile robot in real time, and the robot can issue a "go over there" speech command.

Multimodal Integration of Audio-Visual information for speech recognition is a cross-disciplinary subject that involves processing technique such as image retrieval, proactive information retrieval, and speech recognition and natural language processing and so on. Inspired by lip reading, the correlation between vision and speech information plays an important part in

speech recognition and large numbers of double modal system are applied to speech recognition area. Since speech movement can be dissolved into lip movements and voice spectrum, the correlation between lip movement and speech make this area a natural experiment ground for designing multimodal system and applying multimodal fusion theories. In [18], Nakamura first attempt to propose a multimodal system structure for integrating audio speech and visual face information together for synchronized speech recognition.



(a)



(b)

Fig. 15 (a) Multimodal pen and voice real estate input system on a hand-held PC [13]; (b) The StarWeb system architecture proposed in [30].

Papandreou and Katsamanis et al. [21] developed a visual speech synthesis and fusion system. Geometrical feature of the speaker's mouth is tracked by applying an active appearance modal (AAM) to track feature points of speaker's face. Instead of using mouth shape as complementary information for assisting speech recognition, Chan et al. [44] proposed a method of speech

recognition using a multimodal system combining both acoustic and multichannel facial EMG information. With additional five channels of facial EMG data, Chan was able to classify English words from “zero” to “nine” articulated by two subjects with an accuracy of 85% and 87% when respectively, the system can maintain higher recognition rates in noisy environments compared to a single modal acoustic classifier.

3.4 Virtual reality (VR) and augmented reality (AR) interaction

Augmented reality (AR) is a computer vision technique that uses computer-generated imagery to view a physical real-world environment; normally the elements on the original image are augmented so that the modified image can influence user’s perception of reality. Since Richard Bolt’s pioneering “Put-That-There” system [10], researchers begin to realize that multimodal human’s gestures and speech recognition based communicating can lead to more powerful and natural human–machine interfaces than single modalities. Dominguez, et al. proposed a multimodal wearable computer interface controlled by finger movements and audio control commands with an augmented reality head-mounted camera [22].

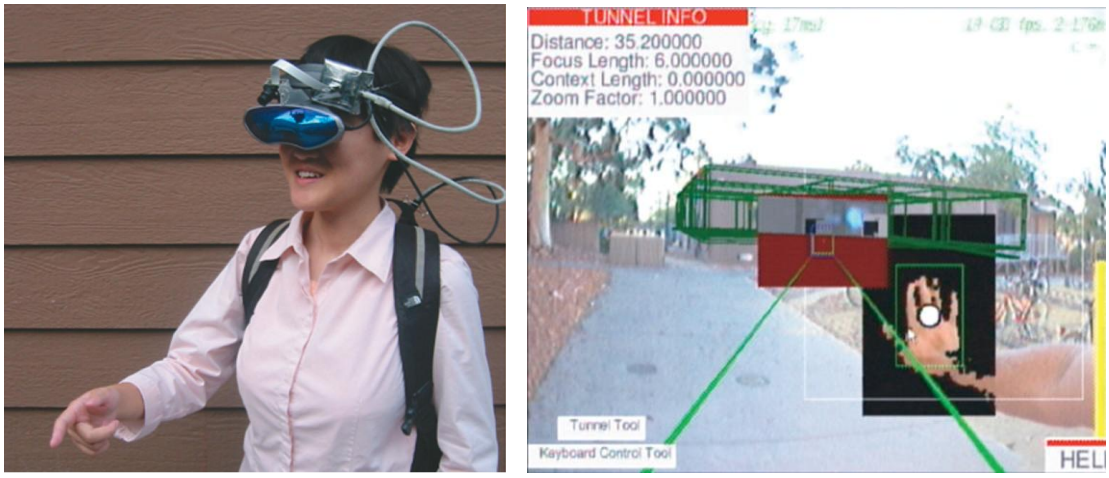
A vision based finger tracking algorithms including color segmentation, fingertip shape analysis, and perturbation model learning are applied to track the trail of finger movements. An object of interest in live images can be divided by finger movements by the user encircling the object with finger movement. IBM’s ViaVoice speech recognition engine are applied to resolve audio control commands such as “clear”, “reset”, “snap” and “reset” which are integrated to enable and disable the finger tips tracking during objects of interest encircling process, and extract the encircled image area from live environment image.

A multimodal gesture and speech controlled interface, namely Masterpiece, are proposed in [23]. The interface can help user generate and manipulate simple 3D objects through the user interacting with a virtual 3D environment in a large screen display. The user’s pointing direction are estimated by tracking user head and hands position in 3D space with the image from a stereo camera, and thereafter projecting tracked head-hand axis to the screen. With a sketching based strategy, five gestural control commands, i.e. “Pointer control”, “Selection”, “Translation”, “Rotation” and “Scaling”, are used for editing the position and the size of the selected object. Eight speech editing commands can are recognized by classifying Mel frequency cepstrum coefficients (MFCC) of voice stream with hidden Markov models. The performance of the interface is evaluated with the user accomplishing a task using the other two interfaces which are CyberGrasp haptic glove and a 3D mouse pointer separately. The result shows that compared with the other two unimodal interfaces, Masterpiece interface has better performance in terms of user immersion, usability, 3D manipulation efficiency, and device intrusiveness.

Fu, et al. constructed virtual 3-D avatar models and interact with the modal using head movements and speech [24]. A head tracker is developed to track user head movements such as rolling, tilting, yawing, scaling, horizontal, and vertical motions, and therefore generate avatar animation head modal which can mimic user head gestures. The lip movement of the user during speaking is modeled by synthesized speech data and analyzing phoneme factors in recognized speech sentence. The recognized sentences are classed into fourteen phoneme groups and for

each group an animated mouth viseme is visualized. Both a male and female avatar modal has been created for commercial kiosk applications.

Kolsch, et al. used a combination of hand gestures, voice, head pose and track ball input to operate an outdoor wearable augmented reality system [26]. As shown in Fig. 16 (a), the head-worn device has a texture and color based Hand Vu computer-vision module to recognize hand gestures and Panasonic Speech Technology Laboratory's speech recognition library ASRlib are used for speech and audio command recognition purpose. They concluded that multimodal user interfaces can well satisfy the diverse input needs of demanding application interfaces on a wearable computers interface.



(a) (b)

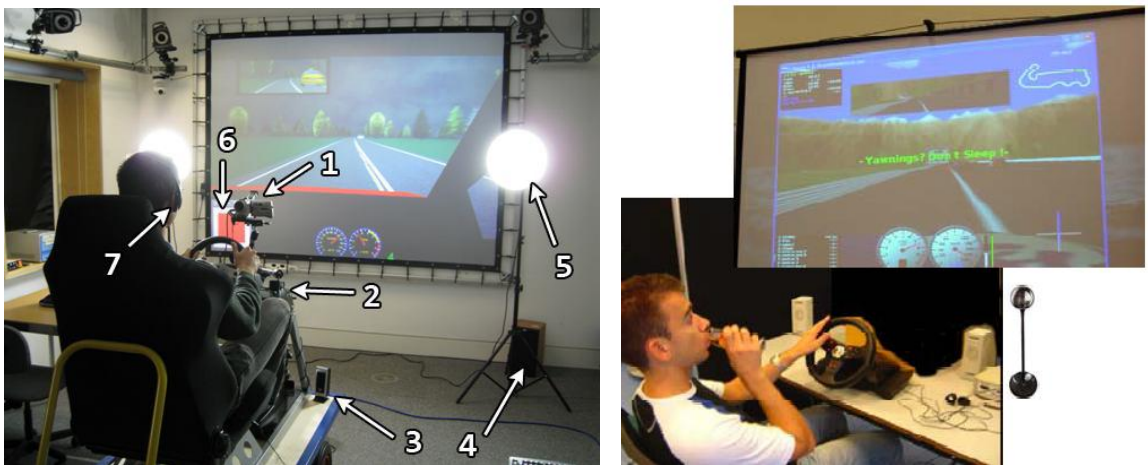
Fig 16 (a) a user wearing a headset multimodal augmented reality system. (b) Augmented reality interface interacted with head gesture [26].

3.5 Driver assisted multimodal interface

Bergasa et al. developed infrared (IR) image based system for monitoring a driver's vigilance when driving [40]. Multimodal facial characters including Percent eye closure (PERCLOS), eye closure duration, blink frequency, nodding frequency, face position, and fixed gaze are calculated and combined using a fuzzy classifier to infer the level of inattentiveness of the driver. Bergasa test the system with different driver image sequences recorded in night and day driving conditions in a motorway and with different users and the result of user drowsiness and inattention detection shows that the fusion of multimodal parameters generate a more robust and accurate result than by using a single parameter.

Similar to Bergasa's work, Ji, et al. combined visual cues of driver from an IR image to infer a fatigue lever of the driver [34]. Multimodal visual cues such as eyelid movement, gaze movement, head movement, and facial expressions are extracted from the driver using various computer vision algorithms; Bayesian networks (BN) models is applied for fusing different visual cues and contextual information to deduce the level of fatigue in real time. Sezgin, et al. developed a driver-vehicle interaction interface (as shown in Fig.17 (a)) using multimodal speech and facial expression information [20]. Benoit, et al. proposed a multimodal driving

simulator shown in Fig. 17(b) [19]. The simulator system capturing and interpreting the driver's focus of attention and fatigue state based on analyzing facial expression, head movement, eye movement from video data acquired from a camera. A Fuzzy Expert System formed by data from 30 drowsy drivers is applied to synthesize and analysis drowsiness factors of current driver using factors such as blinks duration, blink rate and blink interval. Considering the driver cannot share visual attention during driving, head gestures, hand gesture and driver speech are used to control information systems such as radio or telephone and navigation systems on board a car. The interaction interface shows a good performance on both usability and safety since drivers can keep their eyesight on the road during interaction.



(a) (b)

Fig.17 Multimodal driver simulator in [20] (Left) and [19] (Right)

3.6 Multimodal interface for disabled and elderly people

Ju, et al. introduced a vision based intelligent wheelchair system developed for people with quadriplegic [27]. Human mouth shape and head gesture are applied to control a powered wheelchair in both indoor and outdoor environments. The inclination of the head is detected by Adaboost face detector and mouth shape is extracted from edge information of face images. Four control movements: mouth open, mouth close, and head left incline and head right incline are assigned to four control commands, “Go”, “Stop”, “Turn Left” and “Turn Right”, respectively. The system proved that the good recognition rate of user's intention from 34 normal subjects is achieved, and the vision based system can adapt to complicated background and varying illumination conditions.

Li, et al. proposed a bimodal wheelchair control approach by integrating vision and speech controls [50]. Matsumoto, et al. applied the recognition of head motion and eye gaze onto a locomotive wheelchair system [51]. The head motion and eye gaze information are deduced from corners of the eyes and the mouth selected by mouse beforehand. The chair can navigate thorough indoor and outdoor environments according to the direction where the user is looking at, the head movements including nodding and shaking heads are recognized to activate and stop the gaze guided control. Besides computer vision based human machine interface listed above, Ferreira, et al. proposed an HMI structure to control a robotic wheelchair by scalp EMG and

EEG signals [35]. Both eye blinking movements detected in EMG signal patterns and eye close movements recognized from EEM signal patterns are used as command movements to control a mobile wheelchair through an onboard PDA.

4. Discussion and Conclusion

Over the past few years, a large number of human machine interfaces has been developed by combining various user inputs such as speech, pen, touch, hand/head gestures, eye gaze and body movements, etc. And many research issues have been addressed, including facial expression recognition, human emotion analysis, speech recognition/synthesis, human computer interaction, virtual reality and augmented reality interaction, etc. As a result, the development of multimodal HCI systems (MMHCI) becomes a central issue for robotics researchers and scientists, and has many diversified real-world applications, in particular to help elderly and disabled people.

The current applications have paved a way for the future, and will finally result into a universal HMI system that can be customized and inclusive to be fit for different individual users with complicated demands. Although the single modality interaction methods are improved, multi modal interaction methods are on their way of bettering themselves. This mutual benefit and promoting circle can bring a grand and unprecedented advancement era for human machine interaction and can greatly changed our way of life and the way we communicating with machines. Especially, disabled and elderly people will get the great benefit from the applications of multimodal HMI in their daily life.

ACKNOWLEDGMENTS: This research has been financially supported by EU Interreg IV A2 Mers Seas Zeeën Cross-border Cooperation Programme – SYSIASS: Autonomous and Intelligent Healthcare System. More details can be found from the project website http://www.sysiass.eu/. We would also like to thank Robin Dowling for the technical support on wheelchair hardware.

References

[1] R. Sharma, V. Pavlovic and T. Huang, “Toward multimodal human-computer interface,” Proceedings of the IEEE, Vol.86, No. 5, pp. 853-869, 1998.

[2] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey,” Computer Vision and Image Understanding, Vol. 108, No. 1-2, pp. 116-134, October 2007.

[3] C. Busso and S. Narayanan, “Interrelation between speech and facial gestures in emotional utterances: a single subject study,” IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp. 2331-2347, November 2007.

[4] L. Kessous, G. Castellano and G. Caridakis, “Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis,” Journal on Multimodal User Interfaces, Vol. 3, No. 1-2, March 2010.

[5] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, G. Volpe, “Toward real-time multimodal processing: EyesWeb 4,” Proceedings of AISB 2004 Convention: Motion, Emotion and Cognition, Leeds, UK, March 2004.

[6] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers and, T.S. Huang, "Authentic facial expression analysis," Image and Vision Computing, Vol. 25, No. 12, December 2007.

[7] T. Felzer and B. Freisleben, "HaWCoS: The 'Hands-free' Wheelchair Control System," Proceedings of International ACM SIGACCESS Conference on Computers and Accessibility, pp. 127–134, ACM Press, 2002.

[8] R. Barea, L. Boquete, M. Mazo, E. Lopez, "System for assisted mobility using eye movements based on electrooculography," IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol.10, No.4, pp. 209-218, Dec 2002.

[9] X. Huo, J. Wang and M. Ghovanloo, "A magneto-inductive sensor based wireless tongue-computer interface", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 16 No. 5, pp. 497-504, 2008.

[10] R.A. Bolt, "'Put-that-there': Voice and gesture at the graphics interface," Proceedings of the 7th annual conference on Computer graphics and interactive techniques, Vol. 14, No. 3. pp. 262-270, July 1980.

[11] A. Sears, M. Young and J. Feng (2008). Physical Disabilities and Computing Technologies: An Analysis of Impairments. In: A. Sears and J. Jacko (Eds) The Human-Computer Interaction Handbook (2nd Edition). Mahwah, NJ: Lawrence Erlbaum and Associates, pp. 829-852.

[12] X. Huo and M. Ghovanloo, "Using Unconstrained Tongue Motion as an Alternative Control Mechanism for Wheeled Mobility," IEEE Transactions on Biomedical Engineering, Vol.56, No.6, pp.1719-1726, June 2009.

[13] S. Oviatt, "User-Centered Modeling for Spoken Language and Multimodal Interfaces," IEEE MultiMedia, Vol. 3, No. 4, pp. 26-35, Dec. 1996.

[14] I. Iturrate, J. M. Antelis, A. Kübler and J. Minguez, "A noninvasive brain-actuated wheelchair based on aP300 neurophysiological protocol and automated navigation," IEEE Transactions on Robotics, Vol.25 No.3, pp.614-627, June 2009.

[15] "Emotiv Systems Homepage". www.Emotiv.com

[16] W. A. König, R. Rädle and H. Reiterer, "Interactive design of multimodal user interfaces," Journal on Multimodal User Interfaces, Vol. 3, No. 3, April, 2010.

[17] B. Dumas, D. Lalanne and S. Oviatt, "Multimodal Interfaces: A Survey of Principles, Models and Frameworks," Human Machine Interaction: Research Results of the MMI Program, Springer-Verlag, pp. 3-26, 2009.

[18] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," IEEE Transactions on Neural Networks, Vol.13, No.4, pp. 854- 866, Jul 2002.

[19] A. Benoit, L. Bonnaud, A. Caplier, I. Damousis, F. Jourde, J-Y L. Lawson, L. Nigay, M. Serrano, and D. Tzovaras, "Multimodal Signal Processing and Interaction for a Driving Simulator: Component-based Architecture," Journal on Multimodal User Interfaces, Vol. 1, No 1, pp. 49-58, March 2007.

[20] T. M. Sezgin, I. Davies and P. Robinson, "Multimodal Inference for Driver-Vehicle Interaction," Proceedings of the 2009 international conference on Multimodal interfaces, pp. 193-198, 2009.

[21] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive Multimodal Fusion by Uncertainty Compensation With Application to Audiovisual Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing, Vol.17, No.3, pp.423-435, March 2009.

[22] S. M. Dominguez, T. Keaton and A. H.Sayed, "A Robust Finger Tracking Method for Multimodal Wearable Computer Interfacing," IEEE Transactions on Multimedia, Vol.8, No.5, pp.956-972, Oct. 2006.

[23] K. Moustakas, M.G. Strintzis, D. Tzovaras, S. Carbini, O. Bernier, J.E. Viallet, S. Raidt, M. Mancas, M. Dimiccoli, E. Yagci, S. Balci and E.I. Leon, "Masterpiece: physical interaction and 3D content-based search in VR applications," IEEE Multimedia, Vol.13, No.3, pp.92-100, 2006.

[24] Y. Fu, R. Li, T.S. Huang and M.Danielsen, "Real-Time Multimodal Human–Avatar Interaction," IEEE Transactions on Circuits and Systems for Video Technology, Vol.18, No.4, pp.467-477, April 2008.

[25] S. Pastoor, J. Liu and S. Renault, "An experimental multimedia system allowing 3-D visualization and eye-controlled interaction without user-worn devices," IEEE Transactions on Multimedia, Vol.1, No.1, pp.41-52, Mar 1999.

[26] M. Kolsch, R. Bane, T. Hollerer and M. Turk, "Multimodal interaction with a wearable augmented reality system," IEEE Computer Graphics and Applications, Vol.26, No.3, pp.62-71, May-June 2006.

[27] J. S. Ju, Y. Shin and E. Y. Kim, "Vision based interface system for hands free control of an Intelligent Wheelchair," Journal of neuroengineering and rehabilitation, Vol. 6, 2009.

[28] K.-H. Kim, K.-H. Kim, J.-S. Kim, W. Son and S.-Y. Lee, "A Biosignal-Based Human Interface Controlling a Power-Wheelchair for People with Motor Disabilities," ETRI JOURNAL, Vol.28, No.1, pp. 111-114, 2006.

[29] H. Meng, S. Oviatt, G. Potamianos and G. Rigoll, "Introduction to the Special Issue on Multimodal Processing in Speech-Based Interactions," IEEE Transactions on Audio, Speech, and Language Processing, Vol.17, No.3, pp.409-410, March 2009.

[30] D. Sonntag, R. Engel, G. Herzog, A. Pfalzgraf, N. Pfleger, M. Romanelli and N. Reithinger. SmartWeb Handheld: Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In Huang, Th., Nijholt, A., Pantic, M., Pentland, A. (eds.): Artificial Intelligence for Human Computing. Springer, Heidelberg, pp. 272 – 295, 2007.

[31] M. Pantic and L.J.M. Rothkrantz, "Toward an affect-sensitive multimodal human–computer interaction," Proceedings of the IEEE 91 (9) (2003), pp. 1370–1390.

[32] K.H. Kim, S. W. Bang and S. R. Kim (2004). Emotion recognition system using short-term monitoring of physiological signals, Medical & Biological Engineering & Computing, Vol. 42, 419–427.

[33] P. Ekman, Emotion in the human face. Cambridge University Press, 1982.

[34] Q. Ji, Z. Zhu, P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," IEEE Transactions on Vehicular Technology, Vol.53, No.4, pp. 1052- 1068, July 2004.

[35] A. Ferreira, R. L. Silva, W C Celeste, T. F. Bastos Filho and M. Sarcinelli Filho, "Human–machine interface based on muscular and brain signals applied to a robotic wheelchair," Journal of Physics: Conference Series, volume 90, 012094, 2007.

[36] Z. Zeng, J. Tu, B.M. Pianfetti and T.S. Huang, "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM," IEEE Transactions on Multimedia, Vol. 10, No. 4, pp.570-577, June 2008.

[37] E. Mower, M.J. Mataric and S. Narayanan, "Human Perception of Audio-Visual Synthetic Character Emotion Expression in the Presence of Ambiguous and Conflicting Information," IEEE Transactions on Multimedia, Vol.11, No.5, pp.843-855, Aug. 2009.

[38] R. O'Grady, C.J. Cohen, G. Beach and G. Moody, "NaviGaze: enabling access to digital media for the profoundly disabled," Proceedings of 33rd Applied Imagery Pattern Recognition Workshop, pp. 211- 216, Oct. 2004.

[39] M. Song, M. You, N. Li and C. Chen, "A robust multimodal approach for emotion recognition," Neurocomputing, Vol.71 No.10-12, pp. 1913-1920, June, 2008.

[40] L.M. Bergasa, J. Nuevo, M.A. Sotelo, R. Barea and, M.E. Lopez, "Real-time system for monitoring driver vigilance," IEEE Transactions on Intelligent Transportation Systems, Vol.7, No.1, pp.63-77, March 2006.

[41] H. Gunes and M. Piccardi, "Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol.39, No.1, pp.64-84, Feb. 2009.

[42] Z. Obrenovic and D. Starcevic, "Modeling Multimodal Human-Computer Interaction," Computer, Vol.37 No.9, pp.65-72, September 2004.

[43] S. Oviatt, "Advances in robust multimodal interface design," IEEE Computer Graphics and Applications, Vol.23, No.5, pp. 62- 68, Sept.-Oct. 2003.

[44] A.D.C. Chan, K.B. Englehart, B. Hudgins and D.F Lovely, "Multiexpert automatic speech recognition using acoustic and myoelectric signals," IEEE Transactions on Biomedical Engineering, vol.53, no.4, pp.676-685, April 2006.

[45] K. Takahashi, "Remarks on emotion recognition from multi-modal bio-potential signals," IEEE International Conference on Industrial Technology, Vol.3, No., pp. 1138- 1143 Vol. 3, Dec. 2004.

[46] A. Savran, K. Ciftci, G. Chanel, J. C. Mota, L. H. Viet, B. Sankur, L. Akarun, A. Caplier and M. Rombaut, "Emotion Detection in the Loop from Brain Signals and Facial Images," Proceedings of the eNTERFACE 2006, Dubrovnik, Croatia, July, 2006.

[47] D. Perzanowski, A.C. Schultz, W. Adams, E. Marsh and M. Bugajska, "Building a multimodal human-robot interface," IEEE Intelligent Systems, Vol.16, No.1, pp. 16- 21, Jan-Feb 2001.

[48] "SmartNav Homepage" , http://www.naturalpoint.com/smartnav/

[49] B. Rebsamen, E. Burdet, C. Guan, H. Zhang, C. L. Teo, Q. Zeng, C. Laugier, Jr. Ang and H. Marcelo, "Controlling a Wheelchair Indoors Using Thought," IEEE Intelligent Systems, Vol. 22, No. 2, pp. 18–24, 2007.

[50] X. Li, T. Tan and X. Zhao, Multi-modal Navigation for Interactive Wheelchair, Advances in Multimodal Interfaces — ICMI 2000, Vol. 1948/2000, pp. 590-598, Springer Berlin Heidelberg.

[51] Y. Matsumoto, T. Ino and T. Ogasawara, "Development of Intelligent Wheelchair System with Face and Gaze Based Interface," Proceedings of 10th IEEE Int. Workshop on Robot and Human Communication (ROMAN 2001), pp. 262-267, 2001.