ABSTRACT

Hand gestures and speech constitute one of the most important modalities in human-to-human communication. The investigation of this phenomenon has been the focus of research about multimodal interfaces for speech and hand gesture recognition in human computer interaction (HCI) during the past decade. As one of the most popular application areas of HCI, computer games are likely to make use of this technology. Combining speech and gesture in computer games may enhance the quality of entertainment while making game play more enjoyable. This article does not intend to cover exhaustively all the issues related to multimodal interface creation in computer games. We primarily discuss some of the key issues related to multimodal interfaces for speech and hand gesture recognition in computer games and present a survey of current approaches to overcome the problems faced in this area.

1. INTRODUCTION

To date, some of the human communication modalities (e.g. speech) have been extensively investigated. Speech recognition rate has reached a high level of performance in controlled situations. Advent of commercially successful systems for speech recognition allow to apply human voice in HCI. Human movements, especially hand gestures, has also become an important part of HCI in recent years. The advances in gesture recognition allow many practical applications (e.g. [1]). There are studies stating that while interacting with computer systems, people prefer a combination of speech and gestures over speech and gestures alone [2]. Different input modalities can complement each other and allow greater expressiveness than each modality on its own. The modalities can also enhance each other when similar concepts are expressed in many different ways.

An example of early applications of multimodal interfaces is probably Bolt’s ”Put That There” system [3] which processes multimodal speech and manual pointing during object manipulation. Other similar applications derived from this concept includes multimodal systems combining speech and pen input [4] used in mobile environments and for handfree painting on virtual paper [5]. Users are able to operate the system using spoken language and hand gestures in similar ways face-to-face communication by analyzing and interpreting the continuous streams of speech and gesture signals.

Another potential application area of for multimodal surfaces is Virtual Reality (VR) platforms. VR technology is nowadays flourishing with the rapid development of HCI and related technologies. The wide-range of VR applications include training, education, entertainment, engineering, medical operation etc. The lack of natural interfaces is a main problem for using VR technology widely in computer games. In order to solve this problem, more expertise is required in human perception and natural interaction [6].

The conventional interfaces such as keyboard and mouse limit HCI in virtual and augmented reality applications. The use of multimodal interfaces can help people communicate with computer-based systems in a more intuitive way. A multimodal interface in computer games allows users to exercise and control computers with their own choices of interaction.

The conventional interfaces such as keyboard and mouse limit HCI in virtual and augmented reality applications. The use of multimodal interfaces can help people communicate with computer-based systems in a more intuitive way. A multimodal interface in computer games allows users to exercise and control computers with their own choices of interaction. There are many individuals with various abilities who prefer to adopt different modes of communication. Specifically, the users with sensory impairments and others of temporary or permanent disabilities or illnesses would need personalized interfaces while interacting with computers. In this respect, multimodal interfaces have great potential to serve for a comparatively broader range of users than conventional graphical user interfaces (GUIs). For instance, users with physical disabilities (e.g. injuries of knuckle joint) may prefer using speech and hand gestures for controlling or interacting with computer systems to keyboard or mouse as input devices [7]. Some studies [8] suggested that even visual impaired and hearing and speech impaired people can play computer games through integrating haptics, audio and visual signals together.

For hearing and speech impaired users, the visualization of the hand and body movements can give them access to a virtual game environment. Gesture recognition may allow them to manipulate computer games by gestures. Visual impaired users can interact with a virtual environment by hand gestures and audio information and get feedback through force feedback (transforming sound into haptic feedback). Well-designed multimodal interfaces are expected to be more natural and easier to learn and use for people with disabilities, since they offer users freedom to use a combination of different modalities with different input devices.

To build multimodal interfaces for speech and hand gesture
recognition in computer games, the key issue to investigate is how to combine different input signals to achieve maximum joint benefit. One of the most significant challenges facing multimodal interface design is the development of effective frameworks that can support multimodal systems. Detailed knowledge in every part of the framework, such as input devices, algorithms, and information fusion of modalities, are also crucial for its implementation. In the following parts, we will first review related applications of speech and hand gesture multimodal interfaces in computer games and then discuss the related issues about designing multimodal interfaces for computer games.

2. COMPUTER GAMES WITH GESTURE AND SPEECH RECOGNITION

It has been suggested [9] that players act both as a spectator and a director in most computer games. Speech input would certainly improve the latter role by supplying users with more natural control. This improvement would essentially enable more personal high-level, abstract commands [10]. The experiments carried out in [10] suggest that the performance of speech recognition systems was faster and more accurate in some games (such as Age of Empires and the Age of Kings). They also suggest that speech interface on more than one occasion is convenient in resting hands. A system applying speech commands to a Chinese computer game Mahjong is built and evaluated [11]. This game system has been proven to enhance the interaction between the experimental participants and the computers.

Segan et al. [12] applied hand gesture recognition to computer games and fly-bys using a hand gesture input device. A system by Sony observed players making different fighting gestures and translated those into a computer game [13]. In the work built by the team of Synthetic Characters at the MIT Media lab [14], the user takes on the role of chicken that is trying to protect its eggs from a hungry raccoon. In their system the players are allowed to control the character through a set of gestures representing specific behaviors. In an art installation created by N. Tosa [15], users can influence the behavior of two game characters through their hand gestures. Players can interact with computers directly using hand- and arm-gestures without any other additional equipments except cameras. Lee et al. [16] proposed a Personal Digital Assistant(PDA)-based multimodal interface using speech, gesture and touch sensations for network game. These applications lead to a new type of natural interface for next-generation ubiquitous gaming.

An example [17] that benefits from using multimodal interaction within a gaming scenario, is Lionhead Studios®’ “Black & White®”. This game requests the player to cast spells in the form of letters or symbols by moving the mouse, thus removing the need for extensive menu structures which slow the pace and user’s immersion within the game. By using speech and gestures to accomplish certain tasks users can have much richer and more immersive gaming experience. Kang et al. [18] presented a vision-based interface for video games in which gestures are used as commands to control a video game(Quake II). Their experimental results firmly demonstrated the potential for the development of gesture-based interfaces in computer games. Carbin et al. [19] described a verbal and gesture multimodal human-computer interface which allows users interacting with a chess computer game remotely. These attempts greatly encourage the following researchers to exploit more effective and natural multimodal interfaces in computer games. Based on these pioneering research studies, our aim is to investigate if multimodal interaction in computer games can ease the player’s burden and improve their immersion by distributing the tasks to the most appropriate modalities.

3. FRAMEWORKS FOR SPEECH AND GESTURE MULTIMODAL INTERFACES

In human communication, the use of speech and gestures is completely coordinated. Unfortunately, the devices used to interact with computers in multimodal systems have not been designed at all to cooperate. For example, the difference between time responses of different devices can be very large. A speech recognition system may need more time to recognize a word than a gesture recognition system to compute the point coordinates relative to a pointing gesture [20]. This implies that the order of system receiving an information stream is not consistent with the actual chronological order of user’s actions. In this respect, we need to define certain criteria to decide how and at which level speech and gestures can be fused in a multimodal system. Generally fusing information at a lexical level allows for designing generic multimodal interfaces, however, occasionally fusion errors may occur. In contrast, a fusion at a semantic level is capable of exploiting more robust multimodal systems. It is, nevertheless, usually context-dependent and requires resolving semantic conflicts between speech and gestures appropriately [21]. These issues pose a challenge for building multimodal interfaces.

An appropriate framework is one the most important requirements for the design of multimodal systems. So far multimodal interface solutions are mostly related to simple modalities such as a pen, a touch screen and speech. Solutions of choosing a modality including gestures are yet insufficient. The lack of a framework to support speech and gesture multimodal interfaces may limit the application areas.

Basically, a multimodal system framework is required to fuse inputs from subsystems to handle message exchange between users and application systems. At least it should support time stamping of the beginning and end of individual input signals. Because speech and gesture data streams for multimodal systems are supposed to be delivered either sequentially or simultaneously [7]. Time is an essential factor in in-
Computer-based multimodal interfaces first demand for the perception of outside world in order to interpret and fuse information of input speech and hand gesture signals. Numerous sensing technologies have been developed in the past for speech and gesture acquisition. Speech sensing concerns with capturing verbal information from users. In a multimodal interface system for computer games, speech acquisition is both difficult and crucial. In order to obtain high signal-to-noise ratio (SNR) speech signal, one approach is to make users as close to the microphone as possible by using headsets or lavaliere microphones [30]. Another approach is associated with noise cancelation techniques by employing several additional microphones or distributed microphones array. The additional microphones or microphones array are mainly employed to capture background noise signals. This procedure can help reduce noise effect by the sequential denoising algorithms. [23] mentioned an approach resorting to physically directional microphones (e.g. shotgun [31] or parabolic microphones [32]). These approaches generally perform well depending on various application environments. Generally, in relative noisy gaming environments (such as multiplayers playing game in the same room), using headset microphones is relatively better but this requires users to wear additional devices. The disadvantage of parabolic microphones is to constrain users to interact with a system in a fixed location. By contrast, the distributed microphones array are able to capture voice signals adaptively from arbitrary locations in space but are apt to get lower quality speech signals due to echoes in indoor contexts. Parabolic microphones and microphones array also tend to be out of operation if players enjoy game sound while using speech interface to control game. It is a fact that sound effect is an indispensable part of computer games nowadays and most of players also enjoy them while playing games. The mixed game sound and user’s speech commands would certainly decrease the system performance. As a result, headset microphone is probably a better choice than other input devices for applications in computer games.

4. TRACKING TECHNOLOGIES FOR SPEECH AND GESTURE

Fig. 1. Typical information flow for a speech and gesture multimodal interface. (Adapted from [22])

Fig. 2. Facilitator for speech and gesture multimodal interface
users’ hands with special gloves that can capture hand and finger movements by special sensors [33, 34]. Another available approach for designers is to adopt computer vision based technologies. This method is also more natural than other gesture sensing devices. Vision-based approach involves one or more cameras to collect images of the user’s hands. The cameras then send the grabbed images to image processing routines to perform hand gesture recognition [1, 35, 36, 37]. A Gesture Pendant [38] that can be worn on the user’s chest is described. This allows a user to move freely rather than stand in the visual angle of fixed cameras. A vision-based approach is restricted by its precision compared to glove-based acquisition technology which can track movements of each finger independently. The cumbersome gloves, however, tend to be uncomfortable.

At present some commercial gesture acquisition devices are available on the market, such as SDT Data Gloves [39] developed by fifth dimension technologies and Nintendo Wii remote [40] implemented by using accelerometer and optical sensor technology. The commercial viability for gesture sensing gradually inspires the emergence of multimodal interface systems for speech and gesture recognition. Recently the use of multiple technologies simultaneously for efficient gesture capturing has been described in [41, 42, 43]. These studies may result in a new direction for gesture acquisition. The primary objective of computer games is entertainment. Players are not willing to suffer from discomfort while playing games. Vision-based gesture capturing techniques would likely be more proper for computer game systems. Many related works (such as [24, 18, 19, 44, 45, 46, 47]) adopt cameras as gesture input devices. Data glove technology, however, has been greatly improved with the development of computing and fabrication technology [39]. They are becoming more precise and lighter. A few studies have also applied data gloves in computer games successfully [16, 48].

5. METHODOLOGIES FOR SPEECH AND HAND-GESTURE RECOGNITION

The majority of techniques related to speech signal processing are based on stochastic process methodology due to its nonstable characteristics. Some studies [49] suggested that language models are usually formulated on the basis of Hidden Markov Models (HMMs) which lay a strong foundation for its application in speech signal processing. To date HMM is, undoubtedly, the most popular method for Automatic Speech Recognition (ASR) [50]. However, the HMM loses its performance in the design of robust ASR systems due to the mismatch between training and testing conditions. In the late eighties and early nineties some other alternative approaches have emerged. The most popular one is based on Artificial Neural Networks (ANNs) (e.g. [51, 52]). Some of them deal with ASR problems using its improved methods (e.g. predictive ANNs [52]). Recently others proposed hybrid ANN/HMM approaches [53]. Solera-Urena et al. [50] suggested to use Support Vector Machines (SVMs) to improve the robustness of ASR systems. Kavakli et al. [43] successfully applied SVM to design a speech and gesture based training simulation. These approaches provide a solid foundation for us to study multimodal interfaces including speech as input mode in computer games.

In past decades gesture languages that accompany speech have been shown to be effective tools in multimodal user interfaces [3, 54, 7]. The hand compared with other body parts is the most effective, general-purpose tool considering its flexible functionality in communication and manipulation [1]. Hand gestures performed by different people can be viewed as a realization of a stochastic process similar to speech signals and modeled appropriately [23]. Generally hand gestures have two aspects: the static aspect (hand postures) and the dynamic aspect (dynamic gestures) [55, 1]. Static gestures are characterized by poses or configurations of hands in an image. Dynamic gestures can be defined either as the trajectory of hands or as a series of hand postures in a sequence of images. It is comparatively easier to recognize static gestures than dynamic gestures. Their stochastic characteristics, however, to some extent limit techniques employed to process hand gestures. Similar to speech recognition, approaches based on ANNs and HMMs are still predominating over others in the field of hand gesture recognition. Some studies showed that ANNs handle pen-input hand gestures well [7], whereas other studies [55] suggested that HMM is a good choice for dynamic hand gesture recognition. Hong et al. [56] proposed an approach for 2D gesture recognition which models each gesture to be recognized as a Finite State Machine (FSM) in spatial-temporal space.

Recently, motivated by designing multimodal interfaces, researchers investigate the correlations between speech and hand gestures to improve gesture recognition. The first attempt in [57] demonstrated a concept of improving recognition of coverbal gestures when the visual signal is combined with the pitch information of utterances. They then proposed a framework of gesture and speech co-analysis based on prosodic manifestations in audio and visual channels [58]. This framework uses HMMs for feature co-analysis and a Bayesian network for co-articulation analysis to improve gesture recognition rate.

As we all know, the control of computer games presents dynamic characteristics in most cases (such as moving, running, fighting). It is reasonable to hypothesize that HMM may be a good choice in the design of a multimodal interface for computer games. Because it performs well in both speech and hand gesture recognition, which may simplify the architecture of multimodal interfaces. Some studies have applied HMM to gesture recognition for the design of computer game interfaces (e.g. [59, 8]). Some other studies adopted HMM for speech recognition in computer game design [16, 25]. To the best of our knowledge, a few studies applied HMM both
to speech and gesture recognition in the field of computer games. That is probably because HMM usually requires a large number of training samples compared to other methods. However, the training samples for speech and gesture multimodal interfaces are hard to obtain in most cases. The computational costs are also increasing rapidly with the increase of speech or gesture vocabulary.

6. MULTIMODAL INFORMATION FUSION

As mentioned above integration of multiple modalities have earned more attention than ever before [60]. Practical implementations of multi-modal systems, however, have been slow to emerge. The implementations are normally restricted by lack of understanding about how to combine different input signals to achieve maximum joint benefit.

Unlike human beings who are naturally able to fuse speech signals and hand gestures together and interpret them simultaneously, computer systems with multimodal interfaces present challenges for the integration of complementary modalities to form a highly collaborative blend. Generally, there are three principal levels at which multimodal architectures handle joint processing of input signals [60], mainly involving speech and gesture signals in our scope.

One method is that integrating signals at the signal or data level. Literally two or more signals, generally generated from signal collecting devices, are fused together directly at this level. Combining speech and gestures at this level is a difficult task, because speech is audio signal and gestures normally can be viewed as video signal or electrical signal according to the sensing techniques for gestures. In this respect, speech and gestures thus have different characteristics. At data level, the possibility of information fusion can only be performed adequately for synchronous signals coming from very similar modality inputs (e.g. two webcams collecting the same scene from different angles).

Another method is that integrating signals at the feature level. Generally, closely coupled and synchronized signals such as speech and lip movements are integrated in feature-fusion architecture. High synchronization is a typical problem associated with information fusion at feature level. In the case of speech and hand gestures, input channels may provide asynchronous but complementary information with different natures (e.g. distances and time scales). Therefore, such kind of information integration is subject to failure. At this level a greater volume of data for training are also required which result in the increase of computational processes and cost. Moreover, the corpora for speech and gesture multimodal training have hardly been obtained in current state.

The third method is that integrating information at a semantic level. This type of fusion approach is comparatively easier to implement for speech and gesture multimodal interfaces. The system fusing information at this level extracts semantic information from each individual modes respectively and then integrates them sequentially, instead of directly mixing signals or signal features together. This approach guarantees that information fusion has the advantage of steering clear of the requirement of synchronization issues. It is also capable of easily using existing relatively simple unimodal techniques or some publicly available products. Generally, the individual modes can be trained and maintained separately by using unimodal data and changed according to system requirements without retraining. On one hand, this probably mean that semantic-level fusion has to rely on the quality of previous processing. On the other hand, how to associate the verbal utterances from the speech modality with co-occurring hand gestures observed by the gesture modality is problematic for information fusion at this level.

Understanding how gestures and speech relate in time is critical for the association. In pen-based systems deictic gestures have been demonstrated to happen before the deictic keywords (e.g. "this", "that"), while some other investigations showed that the deictic words occurred during or after the gestures in 97% of the cases for large screen display systems [23]. However, deictic gestures are just a small part of hand gestures people are able to use in computer games. Studies regarding how to associate many other hand gestures to speech are still insufficient. Wu et al. [61] claim that their system named Quickset prefers to integrate a gesture with speech that follows within a 4-s interval, rather than integrating it with preceding speech. They obtained the approximate time interval after about 1539 command tests. A more general conclusion is that both the gesture stroke and the spoken utterance are performed together at more or less regular intervals. These interval turn out to be between 1 and 2 seconds [62].

Eisentein et al. [63] proposes an approach for gesture-speech alignment. In this approach, coverbal gesture recognition can be improved by combining the visual signals with the pitch information of speech. This has been proven to be robust in spoken English in spite of its disfluency. Another investigation [64] tested the hypothesis that some hand or head movements that occur during speech can be timed with respect to prosodic structures of utterances. The results provide firm support for the hypothesis that the accompanying gestures occur in close temporal alignment with the prosodic structures of speech. Correlation between speech prosody and the accompanying gestures show us a potential direction to advance the development of multimodal interfaces effectively.

Semantic-level fusion is nowadays the most common type of fusion in multimodal applications. At the last stage of it, semantic information extracted from speech and gestures is interpreted to an integrated representation. To achieve this interpretation, the following schemes have been proposed: a unification-based integration approach [65], a finite-state parsing method [66], statistical fusion [61], and a graph-based optimization method [67]. We give a brief description of these four schemes as follows.
The unification-based scheme integrates spoken and gestural inputs by unification of typed feature structures representing the semantic contributions of the different modes. This fusion method allows the component modalities to mutually compensate for each others’ errors. However, this approach does not allow for tight-coupling of multimodal parsing. Johnston et al. [66] later proposed a finite-state parsing method which parses, understands, and integrates speech and gestures using a single finite-state machine. Statistical fusion can be viewed as an evolution of standard unification-based approaches. It adds statistical processing to the fusion procedure in unification-based method. The graph-based representation consists of a set of nodes that are connected by a set of edges. Each node encodes the properties of the corresponding input mode including semantic information and temporal information. Each edge represents a set of relations (such as temporal relations and semantic relations in accord with node information) between two modalities. Given these representations, the goal is to find the best match between two graph-based representations.

Although most of the schemes listed above are theoretically feasible, they are rarely used in practical systems. In addition, these approaches are usually context-dependent and constrained within simple contexts. This is simply because that the generality of relationships between speech and gesture, especially across various speaking circumstances and speakers, still remain to be explored. The widely and successfully applied approach among there is unification-based fusion [7]. Carbini et al. [19] also successfully apply this fusion scheme to implement a computer game system for playing chess.

7. CONCLUSION AND FUTURE WORK

This paper presents key issues related to multimodal interfaces for speech and hand gesture recognition in computer games. Generally, most related studies investigating multimodal interfaces in computer games are empirical or heuristic. The performance of the existing multimodal interfaces in computer games are still arguable. Even though successes in algorithm development and computational power greatly help the development of multimodal interfaces in computer games, controlling the motion of characters by speech and hand gestures in computer games is still a difficult task. This field is still young and needs further research to build reliable multimodal systems.

The latest related successful work “Project Natal” [68] about this project may indicate future directions for us to improve the design of multimodal interfaces in computer games. For instance, what happens with two players who have similar voices? For that matter and how do the cameras on the Natal project differentiate between two players?

The listed two questions are actually related to the technical robustness of practical systems which need to be well devised in practice. Other important directions for further research include using some more tangible interfaces such as digital paper and pen, and multi-touch table, surfaces and screens. From the player’s point of view, the usability of multimodal interfaces, naturalness and quality of the interactions should also be well satisfied. Multimodal dialog design and processing in computer games are crucial aspects to satisfy players. This will draw advantage from the the recent and promising research field of emotion recognition in VR. Emotions can significantly change the conveyed message. Affect recognition is most likely to be accurate when it combines multiple modalities. How to implement it in computer game interfaces requires attention. In future, the empirical studies into human-computer interaction in games should be grounded in a general theoretical framework. This will enable us to obtain more general findings to make them become more mainstream, especially since multimodal interfaces are viewed as the most promising avenue for next-generation computer games design.

8. REFERENCES


