Chapter 9: Future of HCI

HCI has contributed much to the advancement of computing and its spread into our everyday living. The prevalent type of interface up to the late 20th century was the so called WIMP (Windows, Icon, Mouse, Pointer) and graphical user interface for the stationary desktop computing environment. This was a huge improvement to its predecessor, the keyboard input command oriented interface. Much innovation has been made on the 2D oriented desktop interface since it was first introduced to the mass in the early 80s. These include ergonomic mouse and keyboard design, hypertext and web interface, user interface toolkits, extension of the Fitt's law, interaction modeling and evaluation methodologies. If you look more closely, the innovation in HCI has always followed or been accompanied with the advancement of the hardware and software platforms. Even though the original "concept" of the mouse and graphical user interface was actually devised in the late 60's by Doug Engelbart [Mouse], it was not until the early 80's when the hardware and software technology (not to mention the possibility of "personal" computing with the much more affordable price tags) was mature enough to accommodate the use of mouse and the GUI.

[Figure 9.1] Keyboard input command oriented interface[1] to the WIMP and GUI based

[1] Apple Computer, http://www.vintage-computer.com/apple_ii_plus.shtml

interface[2](1980 ~ 1990).

This line of thought can give us a good glimpse into the future of HCI based on the fast-changing trends in the computing platforms. Here are four major new computing platforms emerging since the past 10 years:

- **mobile and hand-held platform** (exemplified by the smart phones) which we can carry around to compute and communicate,
- **ubiquitous platform** in which everyday objects are embedded with interactive computing/networking devices and services,
- **natural and immersive computing/sensing/display platform** that provides near-realistic services and experiences, and
- **cloud computing platform** that provides high quality interactive services (based on its heavy duty ultra-server level computing power) with real time response (based on the fast network service).

In the case of the cloud computing platform, the typical user will not interact directly with the system where the application resides (somewhere in the cloud), but through the client computer or device like the everyday desktop computers and mobile devices. Despite the extreme growth in the desktop and even mobile computing power, they, as stand-alone machines, are not usually

[2]Logitech MK330, http://www.logitech.com/ko-kr/product/wireless-combo-mk330?crid=27

sufficient for the high-end interactive and intelligent services such as image recognition, language understanding, context based reasoning, and agent-like behavior. Note that these so called client devices (for the cloud) are getting rather richer in their sensing, display and network capabilities. It is almost like the cloud taking up the role of the "Model" and the client, "View/Controller" where there can be many "View/Controllers" for different types of clients (e.g. desktop, pads, smart phones). This can be viewed as a way to improve the "UX" by providing the high-quality services in real time, and having specialized interaction clients focused on usability, easily deployed (due to their lightness and mobility) in our daily lives. For such an envisioned future, it will be necessary to develop middleware solutions that will manage the seamless connection between the "Model" and one of many possible client "View/Controllers."

On the other hand, hardware wise, we expect the mobile and ubiquitous platforms will accelerate and further drive the integration of embedded computers and sensors/sensor networks into everyday objects (as is the goal of Internet of Things). The touch technology, as the main interaction mode for the mobile and embedded/ubiquitous computers and devices, will become more refined into multi-touch, proximity touch (hovering) and touch-haptic feedback.

The interaction styles of the mobile/embedded vs. natural/realistic/immersive can be understood in terms of people's natural dichotomous desires: one for simple and fast operations in a dynamic environment and the other for rich and experiential one in a more stable relaxed environment. These two desires are in tune with the lifestyle in the coming ages as we become more affluent

and culturally richer. Virtual and mixed reality, multimodal interfaces are in the forefront of the experiential interaction technologies.

Finally, as we have indicated in Chapter 8, in pursuit of the mystical "UX," more interfaces are becoming "affective," calling out to our emotional side. It is difficult define to what constitutes an affective interface. It could be something as simple as an emphasis in the aesthetics. It could mean personalization and adaptive-ness catering to the user's unique and changing tastes and needs. However, the latter still remains a technological challenges it requires intelligent sensing and robust recognition of contexts, user emotions and subtle intent, a very difficult task even for humans themselves. However, the machine intelligence technologies continues makes almost unimaginable leaps as demonstrated well by the recent IBM Watson computer that has beaten a human champion in the quiz show contest [Jeopardy]. In the following sections, we take a closer look at these promising HCI technologies, many of which are in active stage of research.

[Figure 9.2]Four emerging computing platforms and associated HCI technologies to pay attention to in the next 10 years: High quality cloud service and ubiquitous and mobile interaction clients, experiential and natural user interfaces.

9.1 Non-WIMP / Natural / Multimodal Interfaces

You will remember, in Chapter 4, when we studied the process of HCI design, after considering various requirements, user characteristics and operating constraints, not many interface choices were available after all, as we had to consolidate different possible solutions according to the restrictions imposed by the practically limited computing platforms available today (e.g. WIMP for desktop and touch based for smart phones). However, in the coming ages, as we expect many different computing platforms, we are bound to have more choices, including the non-WIMP type of interfaces, almost synonymous to the more natural and multimodal interfaces. One of the main reasons these non-WIMP interface has not made it into the mainstream yet, despite their apparent needs, is because of the lack of robustness and accuracy, or from another perspective, the relatively large amount of required computation to achieve them. However, the situation is changing due to continued technological innovation and the emergence of the cloud computing infrastructure. In the light of this trend, we go over and assess the future of these HCI technologies one by one like the language understanding, gesture recognition, image recognition and multimodal interaction.

9.1.1 Language Understanding

The "talking" computer interface is undoubtedly the holy grail of HCI. Language understanding can be largely divided into two processes. The first is recognizing the individual words, and the second is making sense out of the sentence which is composed of a sequence of recognized

words (usually known as natural language understanding. Surely word recognition (which could be spoken, written or printed) is the prerequisite to the sentence understanding (here we focus only on the spoken word or voice recognition). Voice recognition performance and its practicality are dependent on the target number of words to be recognized, the number of speakers, the level of the noise in the usage environment, a need for any special devices (e.g. noise cancelling microphone). The current state of the art seems to be: (1) over 95% recognition rate (individual words), for (2) at least millions of words and more than 30 languages, (3) in real time (through the high performance cloud), (4) without speaker specific training (by age, gender, dialects), (5) in a mid-level noisy environment (e.g. office with ambient noise of around 30~40 dB), and (6) with the words spoken relatively closely to cheap noise cancelling microphones or software [8]. Such a state of the art seems be quite sufficient for a more widespread presence of voice recognition in our current lives, but it is not so except for special situations of disability support or for operating constraints in which both hands are occupied. One main reason seems to be that the users are less tolerant to the 2~3% of incorrect recognition performance even though humans themselves do not possess the 100% word recognition capability. Another reason might have to do with the "segmentation" problem. Often, voice recognition requires a "mode" during which the input is given in an explicit way, because otherwise it is quite difficult to separate and segregate out the actual voice input from the rest (noise, normal conversation) within the stream of voice. The entrance into this mode will typically involve, albeit simple, additional actions, such as a button push/release. It has been known that users take this to be a

significant nuisance in usage.

One way to perhaps overcome this problem is to rely more on multimodality. To eliminate the segmentation problem, the voice input can be accompanied by certain other modal action such as a gesture/posture and lip movements within certain context so that it is distinguished from noise, other people's speech, or unrelated conversation. We will get to the multimodal integration in the later section.

While isolated word recognition is approaching nearly 100% accuracy rate, when trying to understand a whole sentence, words need to be recognized from a continuous stream of words. By a simple calculation, we can easily see that recognizing a sentence with 5 words, with each word recognition rate of 90%, will yield only of $0.9^5 = 0.59$ success rate. Add the problem of extraction of the meaning of the whole sentence, and now we have an even lower success rate in the correct natural language understanding.

Despite these difficulties, due to its huge potential, great efforts are continually being made to improve the situation. The recent cases of Apple SIRI [5] and IBM Watson [Watson] illustrate the bright future we have with regards to voice/language understanding. Apple SIRI understands continuously spoken words and understands them with higher accuracy by incorporating the contextual knowledge of mobile device usage. IBM Watson showcased a very fast understanding of the quizzes asked in natural language in its bout with the human champion (however the questions were asked in text, not in voice) [7]. While the computer used in the quiz contest was

a near supercomputer level server, IBM is developing a more compact and lighter version specialized to a specific and practical domain such as medical expert systems and IPTV interaction [6]. AT&T provides a similar voice/language understanding architecture for mobile phone usage as shown in Figure 9-3 [attwatson/footnote].

[Figure 9.3] Voice/Language understanding service by the AT&R Watson cloud engine[3].

9.1.2 Gestures

In human communication, gestures play, in many cases unknowingly, a very important role, often by itself or in a supplement fashion to other modes of communication. Consequently, the objective of incorporating gestures into human-computer interaction is only very natural. While there may be many different types of gestures either from the human's perspective (e.g. supplementary pointing vs. symbolic) or from the technological viewpoint (e.g. static posture vs. moving hand gestures), perhaps the most representative one is the movement of the hand(s). Hands/arms are used often for dietic gestures (e.g. pointing) in verbal communication. For the hearing-impaired, the hands are used to express sign language.

To interpret gestures, the gesture, whether it is a static posture or movement of limb(s) in time)

[3]AT&T Labs Research, http://www.research.att.com/articles/featured_stories/2012_07/201207_WATSON_API_announce.html?fbid=RexEym_weSd

must be captured. This is generally called the motion tracking. Motion tracking can involve a variety of sensors and targeted for many different body parts. Here we illustrate the state of the art by looking at the problem of hand tracking first. Good examples of two dimensional hand/finger tracking are the ones using the mouse and touch screen. These technologies are quite mature and highly accurate, helped by the fact that the tracked target (hand/finger) is in direct contact with the devices. In the case of the mouse, the user has to hold the device and this is a source of nuisance, especially if the user is to express 2D gestures rather just using it freely control the position of the cursor. This explains the reason that mouse driven 2D gestures have not seen the light of day all that much so far except in few games [4]. On the other hand, simple 2D gestures, such as swipes and flicks, on the touch screen are quite popular.

With the advent of the ubiquitous and embedded computing, which in many cases will not be able to offer sufficient area/space for 2D touch input, understanding of aerial gestures in the 3D space, which is actually closer to how humans enact gestures in real life and understand by vision, will become important. Tracking of 3D motion of body parts or moving objects is a challenging technological task. The "inside-out" method requires the user to hold (e.g. 3D mouse, Wii-mote) or attach a sensor to the target body part or object (e.g. hand, head), causing cumbersomeness and inconvenience (Figure 9.4). These sensors operate based on variety of underlying mechanisms such as the detecting the phase differences in electro-magnetic waves, inertial dead reckoning with gyros/acceleration sensors, triangulation with ultrasonic waves, etc. The "outside-in" method requires an installation of the sensor in the environment, external to the user body.

Using the camera or depth sensors (e.g. Microsoft Kinect) are examples of the "outside-in" method. Since the user is free of any devices on one's body, the movement and gestures become and feel more natural, comfortable and convenient. With the sensors being remote, the tracking accuracy is relatively lower than the "inside-out" methods.

[Figure 9.4] Examples of "inside-out" type (hand-held) of sensors (3D mouse[4]) for 3D motion tracking and interaction[5].

However, in recent years, camera based tracking has become a very attractive solution because innovations in the computer vision technologies and algorithms (e.g. improved accuracy and faster speed), lowered cost and its ubiquity (virtually all smart phones, desktops, laptops and even smart TVs are equipped with very good cameras), ever-improving processing power (e.g. CPU, GPU, multimedia processing chips), availability of standard and free computer vision/object recognition/motion tracking libraries (OpenCV[6], OpenNI[7]) and the ease of their programming (Processing language [processing]).

There still exist some restrictions. Performance of camera based tracking is susceptible to the

---

[4]logitech 3D mouse, 3D Mouse & Head Tracker Technical Reference Manual, www.logitech.com

[5]SpaceControl 3D mausmit Ball, http://www.spacecontrol-industries.de/14.html?&L=2

Valentinheun 6D, http://www.valentinheun.com/portfolio/6d/

Novint Technologies Falcon, http://www.novint.com

[6]Open Source Computer Vision (OpenCV), http://opencv.org/

[7]OpenNI the standard framework for 3D sensing, http://www.openni.org/

environment lighting condition. For highly robust tracking, markers (e.g. passive objects that are easily and robustly detectable by computer vision algorithms) are used, which makes the situation similar to using the "inside-out" method. Examples of markers include objects with high contrast geometric patterns, colored objects, and infra-red LEDs.

[Figure 9.5] Camera based motion tracking examples (for face, hand, marker, and whole body).

The inexpensive depth sensors introduced in the market recently has revolutionalized the applicability, robustness and practicality of the "outside-in" gesture and motion based interaction. For example, Microsoft Xbox game platform uses both a color camera and a depth sensor (originally developed by PrimeSense) [primesense] and can track the whole body skeletonal motion (e.g. up to more than 10 joints) of multiple users without any devices worn on the body [openni]. It was originally intended for motion based whole body games, and now its application has been extended to environment reconstruction (i.e. scanning the environment objects to derive computer models), motion capture, and many others. The smaller miniaturized (with comparable resolution and performance) version for mobile devices is expected or already have been developed [ref].

[Figure 9.6] Whole body skeletonal tracking using the Kinect depth sensor (left) and its application to motion based games (right).

[Figure 9.7] Prototype miniature depth sensor mountable on mobile devices [ref].

With all this said, it seems the major hurdle has been eliminated on our road to more widespread use of motion based interaction. There still remains one more problem, which is again the same "segmentation" problem, that was associated with the voice recognition. Similarly, it is a difficult problem to segment out the meaningful gestures out of the continuous motion tracking data. Figure 9.8 illustrates the problem and its difficulty. Again, many current motion gesture systems rely on operating in a particular mode (e.g. applying the gesture while pressing a button, or being in a particular state). However, this defeats the very purpose of the bare hand and truly "outside-in" sensing. Plus, as already stated, this additional step in the interaction, having to enter the "gesture input" mode, lowers the usability dramatically. Innovative algorithms such as those based on the concept of "sliding windows" (continuously monitoring a fixed or variable length of motion stream for existence of a meaningful gesture) may be able to solve this problem.

[Figure 9.8] Three major steps in gesture recognition: (1) motion tracking, (2) segmentation (using the monitoring through the "sliding window" into the tracking data stream, and (3) recognition given the tracking data segment.

The segmentation problem is more serious for gesture recognition because in the case of voice recognition, in many cases, the background noise may be low and the detectable spoken inputs are intermittent, meaning the voice recognition mode can be automatically activated by "sound" activation (e.g. sound intensity is greater than some theshold). Touch gesture is the same. It is natural to expect touches, in most cases, only when a command is actually needed. Thus a touch simply signals the start of the gesture input mode. As for 3D motion gestures, users usually continually move and only part of it may be gestural commands that need to be extracted. Again as we have indicated, multimodal interaction can partly solve this problem. Finally, in terms of usage, while motion based interaction may be experiential and realistic, one must remember it is easily tiring. This important aspect must be taken into consideration when conducting the HCI design for the given application.

So far, we have mostly explained our point using hand or bodily motion and potential difficulties in its detection and recognition. Another special case of using gestures is that of using fingers. Due to the current resolution of the sensors and the relative size of fingers against the larger human body, it is not very easy to detect the subtle articulation of the fingers. Again, with the

current trends in new sensor development and declining cost, this will not be such a big problem in the near future. Depth sensors specialized for finger tracking are already appearing in the market (e.g. Leap Motion [3]). In fact, finger tracking used to be handled in the "inside-out" fashion by employing glove-type sensors. Wearing gloves and interacting with it turned out to be very cumbersome with low usability. More importantly, regardless of the type of sensors used, it must be questioned how valuable finger based interaction might be in improving the UX. In real life, fingers are mostly used for grasping, and rarely as gestures (except for the special case of sign language). Even finger touch gestures (for touch screen interaction) are not that many (e.g. swipe, flick, pinch). It may be possible to define many finger based gestures once detailed finger tracking is technologically feasible, but its utility is questionable. Electromyogram (EMG) sensors are newly used to recognize motion gestures. EMG sensors can approximately detect the amount of joint movement. Figure 9.10 shows a wrist-band type of EMG sensor with which a user is taking a "gun triggering" gesture in a first person shooting game.

[Figure 9.9] Finger based interaction using the Leap Motion [3].

[Figure 9.10] Wrist band type of EMG sensor for simple gesture recognition[8].

[8] http://www.thalmic.com

9.1.3 Image Recognition and Understanding

Image recognition or understanding is perhaps a lesser used technology in HCI, especially for rapid paced and highly frequent interaction in which the use of mouse/touch/voice input is more common. For instance, the most typical use for face recognition might be for initial authentication (as part of a log-in procedure). Object image recognition might be used in an information search process as an alternative to the usual keyword text driven, e.g. when the name of the object is not known or when it happens to be more convenient to take the photo than typing or voicing in the input. Rather, the underlying technology of image recognition is more meaningful as an important part of object motion tracking (e.g. face/eye recognition for gaze tracking, human body recognition for skeleton tracking, object/marker recognition for visual augmentation and spatial registration).

Lately, image understanding has become even more important as the core technology for mixed and augmented reality (MAR) has attracted much interest lately. MAR is the technology for augmenting our environment with useful information (Figure 9.11). With the spread of smart phones equipped with high resolution camera, GPUs, light and fashionable see-through projection glasses, and not to mention near 2 GHz processing power, MAR has started to find its way into main stream usage and may soon revolutionalize the way we interact with the everyday objects. Moreover with the cloud infrastructure, the MAR service can become even more robust and high quality. Finally, image recognition can also assume a very important supplementary role in

multimodal interaction. It can be used to extract affect properties (e.g. facial expression), disambiguation of spoken words (e.g. dietic gestures (Figure 9.12) and lip movements).

[Figure 9.11] Image recognition for (a) face, (b) object/marker[9], (c) hand and their applications for motion tracking and augmented reality.

[Figure 9.12] Bolt's pioneering "Put that there" system. The target object of interest is identified from voice and deitic gestures [1].

[Figure 9.13] Applying image understating to information search and augmentation on a wearable display device (Google Glass[10]).

9.1.4 Multimodal Interaction

Throughout this chapter, I have eluded to the need for multimodal interaction in many occasions. Even though machine recognition rates in most modalities are approaching 100% (with the help of the cloud-client platform), the usability is not as high as we expect for various operational restrictions (e.g. ambient noise level, camera/sensor field of view, interaction distance, line of sight,

[9]Sony Smart AR, http://www.sony.net/SonyInfo/News/Press/201105/11-058E/

[10]Google glass, https://plus.google.com/

etc.). To reiterate, this is where the multimodal interaction be of great help. In this vein, multimodal interaction has been an active field of research in academia since the first pioneering system, called the "Put that there," developed by Bolt et al. at MIT in the early 80s [1]. Since then, various ways of combining multiple modalities for effective interaction have been devised. Although we have already outlined them in Chapter 3, we list them again here.

- **Composed -** In this scheme, for a set of subtasks (which together satisfies a larger task), we assign the most appropriate modality to each task. Thus each modality takes up different roles in the interaction. "Put that there" system was one such example, where the voice was used to understand the action command (verb) and the dietic gesture to identify the target object (pronoun). By "most appropriate" we assume and mean that certain modality is most fitting and natural for certain type of action. For instance in a game application, it can be argued that various settings (e.g. selection of the character, weapon, sound options, etc.) can be accomplished with voice or touch interaction for the highest efficiency, while the game itself using action gestures for the experience. Note that multimodal interaction does not necessarily mean that different modal interactions occur simultaneously.

[Figure 9.14] Multimodal interaction in games (a) using the buttons for setting selection, and (b) action gestures for the game play itself[11].

[11]Nintendo Wii, http://www.nintendo.com/wiiu/features/
Macrosoftxbox, http://marketplace.xbox.com/en-US/Product/Kinect-Sesame-Street-TV/

- **Alternative –** In this scheme, as the name suggests, multiple modal interaction techniques are used for the same subtask independently. The choice is made purely by user preference or by the operational situation. When dialing in a regular situation, one might use the touch interaction, while during driving, voice interaction can be used instead. This way the usability is improved by catering to the user's preferences and needs.

[Figure 9.15] Alternative multimodal interfaces in the vehicle navigation systems (touch and voice)[12].

- **Redundant –** In the "Redundant" scheme, many modalities are used together (simultaneously or not) for the same task (input or output). As an interaction method, it makes the conveying of the intent or information much more robust, by combining those of the individual. For instance, an indication of an incoming phone call can use all three modalities, the visual, aural and tactile (vibration). This way, the user has a less change of missing the phone call.

[Figure 9.16] Redundant multimodal output for an incoming phone call using the visual, aural and tactile.

[12]Finedrive, http://www.fine-drive.com/

Another advantage of multimodal interaction is that, to some degree, parallel interaction is possible. We often find people multitask in different modalities, e.g. walking, listening to music and texting to someone. The extent of this ability is still a research question. However it seems quite certain that for this to happen (in the effective and meaningful way), the (multi)tasks must be independent of each other. If each modal interaction shares a common resource, it can be difficult to multitask concurrently (e.g. listening to music, interpreting the words and dancing to it).

Thus, designing for multimodal interaction requires careful considerations of things like modality appropriateness (for the task), cognitive resource usage, synchronization (e.g. multiple modalities perceived as one event when temporally synchronized with a short amount of time), balance (e.g. one modality is not relatively dominating over one another) and consistency (e.g. providing consistent information content between simultaneous multimodal input/output).

9.2 Mobile and Hand-held Interaction

It goes without saying that the smart phones have now almost replaced the PCs at least in terms of casual computing and even a big part of business computing. As such the importance of usability and UX for mobile and hand-held interaction is even higher than ever. It is also interesting that the mobile device, as represented by the smart phones, is a focal point to which the two notable future trends are converging together: (1) multimodal interaction (with all the on-mobile sensors and displays) and (2) cloud based services (through the high speed wireless communication).

In this context, more research is needed in the ergonomic aspects of multimodal interaction for the active (e.g. while moving), dynamic (e.g. frequently changing operating environment) and multitasking living style. At least one notable trend in the mobile interaction is the "simple and quick" approach (vs. the rich experiential). It is not surprising after all, that people would prefer the simple and quick interfaces in the midst of the modern hectic lifestyle, even for entertainment applications such as games. Many recent successful mobile games are those that are called "casual," in which a single game play session lasts only about a minute with single touch operation and almost no learning required. On the other hand, home-based computing platforms (e.g. game consoles, smart TV, desktop) which would be used in a more relaxed atmosphere are becoming more natural, immersive and experiential.

[Figure 9.17] Two bipolar directions in future interaction style:

(a) "simple and quick" mobile/hand-helds (b) rich and experiential stationary platforms at home[13].

As part of the cloud and to supplement and complement the on-mobile sensors, one particular service to take a note of is the sensor network service, i.e. a network of sensors in the environment collectively providing certain service mediated through the cloud. Sensor networks for example can help the mobile client infer the context of usage (e.g. location/area, lighting

[13]LG Smart TV, http://www.lg.com/tw/smart-tvs

condition, time, no. of people in the vicinity, outdoor/indoor, etc.) and provide UX at the personalized level.

[Figure 9.18] Indoor tracking of mobile devices / user using a Wifi sensor network.

9.3 High end cloud service – Multimodal client interaction

Many interaction technologies require artificial intelligence (AI). After all, recognizing spoken words, sentences, images, and gestures are hallmarks of human intelligence. Advanced AI generally requires large data bases, long off-line learning processes, and often heavy on-line computation (for real time responses). High performance servers coupled with mobile clients that handle the fast input data capture and transfer offer an attractive solution. For example, Qualcomm's Vuforia is cloud based solution for image recognition that can be used for a variety of interactive services such as augmented reality and image based search [vuforia]. To develop an interactive image based service, the developer first registers images of target objects to be recognized in the server ahead of time. These input target images are trained off-line on the server so that they can be recognized well from different viewpoints at different scales and lighting conditions. The mobile application captures an arbitrary image and sends it to the server built with references to the target images of interests. The recognition computation is carried out on the server with the results sent back to the mobile application for further processing (e.g. augmentation on the screen) all in real time (Figure 9.19).

[Figure 9.19] A cloud based image recognition service from Qualcomm's Vuforia[14].

Such a division of computational labor is reminiscent of the old time shared computing scheme. The implication is that such a framework is readily applicable for a variety of HCI related computations such as context based reasoning, multimodal integration, user characteristics deduction, large scale and multi-user tracking, usage pattern analysis, client platform adaptation, crowd-sourcing and big data gathering, environment sampling method, etc. Figure 9.20 and 9.21 illustrate the future vision, in which a middleware installed both at the cloud and client mediate the seamless integration between the two. For instance, the client can register itself with the cloud with information of its sensing and display capabilities. Interactions of the applications in the server can be described and coded only in abstract terms and communicated to the client for actual realization based on the known capabilities of the client device. This way, different models and types of devices can use the same cloud applications and services with interaction customized for users and the particular devices.

[Figure 9.20] The middleware between the cloud and interaction client will enable the vision of "one application – many devices" without separate platform specific implementations.

[14]Qualcomm vuforia, https://developer.vuforia.com/resources/dev-guide/getting-started

[Figure 9.21] Middleware architecture for supporting the cloud-client application platform.

9.4 Natural / Immersive / Experiential Interaction

Today's home computing environment is fast changing with the evolution of the television. The "smart" TVs are no different than a high performance computer with a network connection. Moreover as smart TVs in the living room serve the purpose of center of entertainment and they are becoming more and more "high-fidelity," e.g. PC level computing power, more than 42 inch sized screen with UHD resolution and stereoscopy, 5.1 surround sound, sensors (camera, depth sensor, microphone, etc.), and fiber optic / land line network connection. The recent successes of the Microsoft Kinect and Nintendo Wii games attest to this future trend. Thus we can even expect things like haptic sofa, living room table computing, simple olfactory displays. The applications will eventually extend, initially from entertainments, to immersive teleconferencing for home offices, VR based training and education. Critical to such a future vision will be the VR/immersive/natural UI based contents production pipeline, starting with contents authoring tools. Such contents authoring tools with capabilities beyond just game development or multimedia editing are already starting to appear (Figure 9.23).

[Figure 9.22] VR based home entertainment system[15]

[Figure 9.23] An authoring system for immersive and natural UI based contents (Unity3D[16]).

9.5 Mixed and Augmented Reality

Mixed and augmented reality is yet another interaction medium with lots of hype these days. Mixed and augment reality refers to the medium in which the representations of the real and virtual are mixed in some proportion (a term "Virtuality or Mixed Reality Continuum" was coined accordingly [2], Figure 9.24). For example, for contents with mostly the real objects and only a small portion of the virtual is called the "Augmented reality," while the reverse is called the "Augmented virtuality." MAR requires few core technologies, namely object recognition and tracking (Section 9.1.2 and 9.1.3). This is required to spatially register the augmentation right next to the object targeted for augmentation. A looser form of MAR simply augments the information anywhere on the screen. A Google Glass type of application is such an example, where information is projected on the see-through glass at a fixed position (top right corner of the visual field). MAR can improve the usability and UX in interacting with everyday objects because the associated information resides and gets displayed at the same location with instant

[15]Xbox Kinect, http://www.xbox.com/ko-KR/Kinect/School

[16] Unity3d, http://docs.unity3d.com/Documentation/Manual/CustomizingYourWorkspace.html

recognition and access possibilities.

[Figure 9.24] Mixed Reality / Virtuality Continuum [10]. A spectrum is formed according to the relative proportion of the real and virtual representations in the content. At the extreme, there are completely real environment and the purely virtual environment.

9.6 Others

We have briefly looked at several promising technologies and future trends for HCI (in the very subjective view by the author). There are certainly others (which have actually been touted as interfaces of the next generation), which I have not cared advocate due to various perspectives of my own. I will briefly go over them here before wrapping out this book.

- **Wearable computing and interaction –** The smart phone, while almost un-detachable from many users, is not a true form of wearable computer. Wearable computer started from the concept of embedding computers and interaction devices into clothes and things we wear (e.g. hats, belts, shoes, glasses). This integration of "wears" and computing devices have not advanced as much as it was expected during the last decades both technology and usability wise. Even Google Glass concept is facing practical problems such as its weight, power, and privacy issues. It is still questionable whether computer elements need to be interwoven into

our "wears" (except for very special applications).

- **Interaction based on physiological signals** - Much research has been conducted in ways to take advantage our physiological signals such as the brain waves, EMG, ECG, and EEG. It seems very difficult to extract human intention in a useful and major way for HCI from these raw signals. This line of research will probably focus on the HCI for the disabled people.
- **Eye/Gaze tracking and interaction** – HCI is deeply connected with the line of sight. When interacting, we mostly tend to look at the target interaction object. Tracking of the line of sight is many times done by tracking the head direction, rather than the eye balls themselves. In many cases, it is safe to assume that the front head direction is the direction the eyes are looking. There are not too many applications in which the exact eyeball/gaze direction is so important (maybe except for gaze analysis).
- **Facial/Emotion based input** - Affective interfaces based on aesthetic look and feel and more humane output feedback may be important and emerging techniques for improving UX. However, as an input method, it seems we have a long way to go. Input based on user emotion (e.g. facial expression, tone of voice, particular gestures) is very difficult even for humans themselves, and thus would be very difficult to be used as a robust means of interaction.
- **Finger based interaction** – As explained in Section 9.1.2, finger based interaction has been pursued through the use of gloves. Recently depth based sensing allows finger tracking and interaction without the inconvenience of having to wear a glove. Again not too many

applications can be found to apply finger based interaction in the natural way. Contrived finger gestures can be used but they generally incur low usability.

- **3D / Stereoscopic GUI** – Interacting ay manipulating "3D" GUIs (in stereo) have been depicted in many science fiction movies. However tasks that require precise 3D motions are not many. Most system commands are easier with voice or the familiar 2D cursor control.
- **Context based interaction** – Similarly to the case with the emotion based input, inferring "context" in hopes to adapting to the operational situation at hand or to personalize the interface to the user is very difficult. The true user intent is not always clearly manifested explicitly and capturable/interpretable by the sensors and AI.

9.7 Summary

In the future, the competitiveness of software and digital contents will increasingly depend more on the HCI aspect than on the core functionalities. The problem of HCI is becoming more difficult as there are more computing platforms to choose from for different usage situations (e.g. home, office, mobile, sales, vehicles, military, etc.). The HCI design will play a significant role in this sense due to the decline of "standard" desktop platform. HCI will continue to empower humans by providing services, intelligence, knowledge and even power at their finger tips (figuratively speaking).

Reference

[1] Bolt, Richard A. "Put-that-there: Voice and gesture at the Graphics Interface." Proc. of ACM SIGGRAPH, pp. 262-270, (1980).

[2] Milgram, Paul, et al. "Augmented reality: A class of displays on the reality-virtuality continuum." *Photonicsfor Industrial Applications*.International Society for Optics and Photonics, (1995).

[3] Motion Leap. "Leap Motion." Inc. http://www.leapmotion.com, (2013).

[4] Electronic Arts, Inc. "Black and White 2" http://www.ea.com/black-and-white-2-battle-of-the-gods, (2013).

[5] Apple-iOS7-Siri "SIRI", http://www.apple.com/kr/ios/siri/, (2013).

[6] AT&T Labs Research, "AT&T WATSON(SM) Speech Technologies" http://www.research.att.com/projects/WATSON/?fbid=j6ZWSYBnql4, (2013).

[7] IBM Watson "IBM Watson", http://www-03.ibm.com/innovation/us/watson/, (2013).

[8] Wikipedia,"Google Voice Search" http://en.wikipedia.org/wiki/Google_Voice_Search, (2013).

[9] Wikipedia,"Speech Recognition"http://en.wikipedia.org/wiki/Speech_recognition, (2013).

[10] Wikipedia,"Mixed Reality Continuum" http://en.wikipedia.org/wiki/Reality%E2%80%93virtuality_ _continuum,(2013).

[11] T. Lee and T. Höllerer, Handy AR: Markerless Inspection of Augmented Reality Objects Using Fingertip Tracking. In *Proc. IEEE International Symposium on Wearable Computers (ISWC)*, Boston, MA, Oct. 2007.