1

Senses, Perception, and Natural Human-Interfaces for Interactive Displays

Achintya K. Bhowmik
Intel Corporation, USA

1.1 Introduction

Visual displays are now an integral part of a wide variety of electronic devices as the primary human interface to the computing, communications and entertainment systems which have become ubiquitous elements of our daily lives at home, work, or on the go. Whether it be the watches on our wrists, or the mobile phones that we are carrying everywhere with us in our pockets or purses, or the tablets that we are using for surfing the web and consuming multimedia content, or the laptop and desktop computers on which we are getting our work done, or the large-screen television sets at the center of our living rooms, or the presentation projectors in the business meetings, the visual display is the "face" of all these devices to us, the users.

The same applies to a plethora of vertical applications, such as the check-in kiosks at airports, check-out kiosks at retail stores, signages at shopping malls, public displays at museums – the list goes on and on. The wide array of applications and insatiable market demands have fuelled worldwide research and development to advance visual display technologies and products of all form factors in the past decades, ranging from mobile displays to large screens [1–5].

A quick glance at the market size helps us grasp just how pervasive visual displays have become in our lives. In the last five years, according to the display industry analysis firm IHS, the industry shipped nearly 17 billion flat-panel displays [6]. Also, to get a sense of the rate of adoption, the annual shipment of visual displays has grown more than 50% over this period.

In general, an electronic device performs three basic functions: receive instructions from the user, execute certain processing functions according to the instructions and information

Interactive Displays: Natural Human-Interface Technologies, First Edition. Edited by Achintya K. Bhowmik.

received, and present output or results of the processing to the user. As an example, when the author was typing this chapter on his laptop computer, he used a keyboard and a mouse to input the information, the world-processing software application executed on the microprocessor translated the keystrokes and mouse taps into the desired text and format, and the liquid-crystal display of the laptop computer displayed the text on the screen as a real-time visual representation or output. Hence, the display subsystem in such devices already plays a critical role by presenting information to the user – and, until recently, barring some exceptions, the majority of the electronic devices sported a display device whose sole function was just that – to display the visual information.

However, human-computer interaction and user interface paradigms have been undergoing a surge of innovation and rapid evolution in recent years. The ways we interact with computers had already gone through a transformation in the past decades, with the graphical user interfaces that use a mouse and keyboard as input devices replacing the old command-line interfaces that used text-based inputs. We are now witnessing the next revolution with the advent of more natural user interfaces, where the user interacts with the computing devices with touch, gesture, voice, facial expressions, eye gaze, and even thoughts!

Advanced sensors, systems, algorithms, and applications are being developed and demonstrated for natural and engaging interactions, where the computing devices understand the users' expressions and emotions in addition to the intent. These new interface technologies and the ensuing new class of applications present exciting opportunities for the display technology and the consumer electronics industry at large. With the integration of natural user interfaces, the display device morphs from a one-way interface device that merely shows visual content, to a two-way interaction device that also directly receives user inputs and thus enables interactive applications and immersive experiences. The proliferation of touch-screens and touch-optimized interfaces and applications has already brought this transformation to mobile displays, and now the adoption of an extended array of natural interfaces promise to redefine the whole spectrum of displays and systems by making them more interactive.

This book presents a comprehensive treatment of the natural human interface technologies and applications that are enabling the emergence of highly interactive displays and systems. So, what are "interactive displays"? We define them to be the displays that not only show visual information on the screens, but also sense and understand human actions and receive direct user inputs. Equipped with human-like sensing and perception technologies, a "truly" interactive display will "feel" and detect our touch, "hear" and respond to our voice, "see" and recognize our faces and facial expressions, "understand" and interpret our gestural instructions conveyed by the movement of the hands and fingers or other body parts, and even "infer" our intent based on the context.

While these goals may seem rather ambitious, as the examples shown in Figure 1.1 illustrate, systems of various form factors and applications with natural user interaction technologies are already making a large impact in the market by offering easy and intuitive human interfaces. As reviewed throughout the book, significant advances are taking place in natural sensing and inference technologies as well as system integration and application developments, which are expected to bring about new frontiers in interactive displays.

The block diagram shown in Figure 1.2 depicts the generic functional modules and flow of an interactive display system. The interactions between the user and the display system are orchestrated by the interfaces, namely the input and the output blocks shown in the beginning and at the end. The input block consists of sensors that transform the physical stimuli resulting

Figure 1.1 Interactive displays and systems of a wide range of form factors and applications are already gaining a large foothold in the market, and some examples are shown above. The displays in many of these systems assume a new role of direct human-interface device, besides the traditional role of displaying visual information to the user. (See color figure in color plate section).

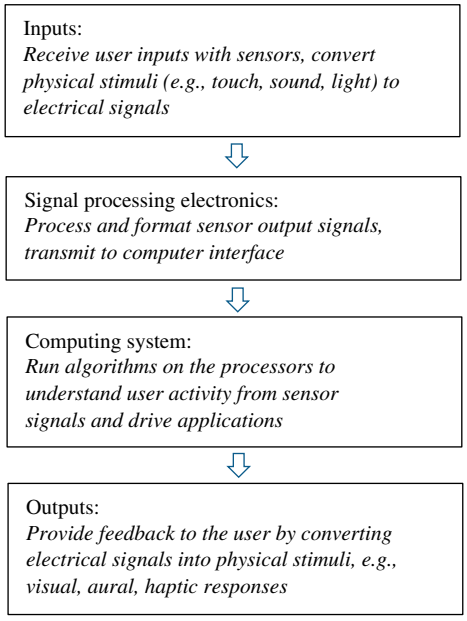

Figure 1.2 Functional block diagram of an interactive display system. The input and the output blocks orchestrate the interactions between the user and the display, while the signal processing and computing functions facilitate these interactions.

from user inputs into electrical signals, while the output block performs the reverse function of providing system responses to user actions in the form of physical stimuli that the users can sense and perceive. The blocks in between perform the necessary signal processing and computing functions to facilitate these interactions.

In this chapter, we first review the basics of human sensing and perception, especially the mechanisms and processes that we deploy in our day-to-day interactions with the physical world. Building on this, we then provide an overview of human-computer interactions

utilizing natural interface technologies based on touch-, voice-, and vision-based sensing and interactions, following a brief review of the most successful legacy interfaces. The subsequent chapters delve into the details of each of these input and interaction modalities, providing in-depth discussion on the fundamentals of the technologies and their applications in human interface schemes, as well as combinations of them towards realizing intuitive multisensory and multimodal interactions. The book concludes with a chapter on the fundamental requirements, technology development status, and outlook towards realizing "true" 3D interactive displays that would provide lifelike immersive interaction experiences.

1.2 Human Senses and Perception

We start with the assertion that the ultimate goal of implementing a human-device interface scheme is to make the interaction experience natural, intuitive, and immersive for the user. While the limitations of the technologies at hand require the designers and engineers to make compromises and settle for a subset of these goals for specific product implementations, we continue to make advances towards realizing this overarching objective.

Let us elaborate on this a little. By *natural*, we mean using our natural faculties for communication and interaction with the devices. We use multisensory and multimodal interface schemes to comprehend our surroundings and communicate with each other in our daily lives, seamlessly combining multiple interaction modalities such as voice, facial expressions, eye gaze, hand and body gestures, touch, smell and taste. The addition of natural interfaces can thus bring lifelike experiences to human-device interactions.

By *intuitive*, we refer to interfaces that require minimal (ideally no) training for the user to engage and interact with the devices, taking advantage of the years of training that we have already gone through in dealing with the world while growing up!

By *immersive*, we allude to an experience where the border between the real world and the virtual world is blurred, with the computers or devices becoming extensions of our body and brain to aid us in accomplishing tasks. This is a tall order, and it will require decades of continued research and development to get closer to these goals. As we endeavor to understand and implement lifelike human interfaces and interaction schemes, it would serve us well to take a look in the mirror, and understand the *human* – after all, that is the first word in the term *human-computer interaction*!

We, the humans, have evolved to be highly interactive beings, aided by a sophisticated set of perceptual sensors and a highly capable brain, including a rich visual perception system, aural and auditory capabilities, touch-sensitive skin and tactile perception, in addition to the chemical sensations of smell and taste via sensors embedded in the nose and the tongue. Well above half of the human brain is dedicated to processing perceptual signals which enable us to understand the space, beings, and objects around us, and interact in contextually-aware natural and intuitive ways.

Let us take a deeper look into three of our perceptual sensors and inference processes – specifically, the eyes and the visual perception process, the ears and the auditory perception process, the skin and the tactile perception process. One reason we limit our discussion to these three modalities of perception is that substantial parts of our interaction with the physical world utilize these mechanisms and, as we will see, imitations of these functions are implementable in electronics devices with the current state-of-the-art technologies in order to design and build highly interactive displays and systems. It would be nice also to implement

smell and taste capabilities in human-computer interactions, but that will have to wait to be the subject matter of another book in the decades to come after further advances are made.

Let us consider the human interfaces with interactive display systems, such as those shown in Figure 1.1, from the neurophysiological perspective. The interaction process can be decomposed into three predominant phases: sensing, perception and recognition, and action. From the viewpoint of the human, the *sensing* process involves collecting the visual output of the display in the form of light waves entering the eyes, the audio output from the speakers in the form of sound waves entering the ears, and a feel of the display surface by touching with the fingertips. These perceptual sensors convert the physical stimuli into neural signals via transduction processes, which are relayed to the cerebral cortex of the brain, where the *perception* of "seeing", "hearing", and "touching" takes place, followed by recognition and understanding.

Based on the results of the perception and recognition processes, we then drive our body parts into *action*. For example, we converge and focus our eyes to the desired elements of the visual content on the display, guide our fingers to touch and activate specific content on the screen, tune our hearing attention to the audio output, sport an appropriate expression on our face, and even articulate a gesture with our fingers and hand.

First, let us review the visual perception process. We will only cover the salient aspects that are relevant to our subsequent discussion on implementing interactive displays, and refer the interested readers to more detailed accounts covered in other publications dedicated to human perception [7, 8]. The human eye is a marvel of evolution, in terms of the sheer complexity of its architecture, the efficacy of its function, and the central role that it plays in our perception of the world in conjunction with the visual cortex in the occipital lobe of the brain. As shown in Figure 1.3, the human eye resembles a camera in some key aspects of its construction, complete with a lens system to focus the incoming light from the scene onto the retina at the back of the eye, which contains the light sensor cells called the photoreceptors. There are two types of photoreceptors in the eye – the color-sensitive "cone" cells and the achromatic "rod" cells that convert the light into neural signals.

How about the *resolution* of this *camera* and the *bandwidth* for communicating with the *processor*? The retina contains a large number of photoreceptors – about eight million cones and about 120 million rods in each eye – and yet the visual system is cleverly designed to signal spatial and temporal changes in the scene, rather than the raw intensity levels detected by the photoreceptors to keep the bandwidth of communication between the eye and the brain down to practical levels.

The visual acuity is sharpest for central vision, when we point our eye to an object and the image is formed within a relatively small area around the optical axis of the eye. This is due to the largest concentration of the cone photoreceptors being in and around the fovea within a small region of the retina, which maps to a disproportionately large area in the visual cortex relative to other parts of the retina. Another important attribute of a camera is the dynamic range of light sensitivity; the human eye spans over ten orders of magnitude, well beyond the capability of our modern-day digital cameras.

While each of our eyes is an elegant *camera*, we have two of them. The human visual system consists of 3D and depth perception capability, with a binocular imaging scheme along with other visual cues such as motion parallax, occlusion, focus, etc., allowing us to navigate and interact with objects in the 3D space with apparent ease. Binocular vision has evolved to be prevalent among most biological systems; recent discovery of fossil records puts it as far back as more than 500 million years ago, to the arthropods of the Early Cambrian era [9]. The advent

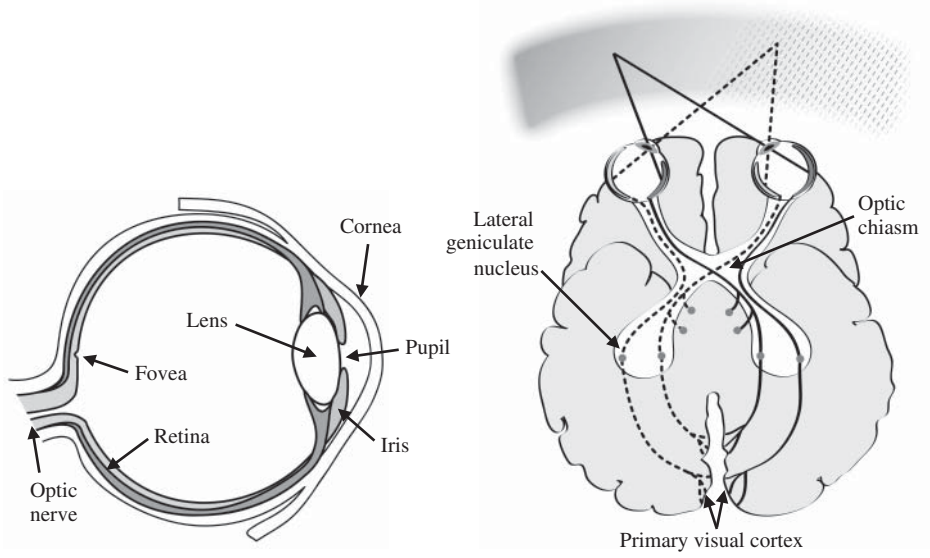

Figure 1.3 Left, anatomy of the human eye. Right, human visual system utilizes a binocular imaging scheme consisting of two eyes. The left visual field is sensed by the right sides of both eyes and is mapped to the right half of the primary receiving area in the visual cortex, whereas the right visual field follows a complementary pathway. The depth of the objects in scene are discerned from the binocular disparity, along with other visual cues such as motion parallax, occlusion, focus, etc.

of powerful visual systems is believed to be a trigger for the Cambrian Explosion of evolution [10]. The partially overlapping visual fields of the two laterally displaced eyes results in a "binocular disparity", where an object is observed to be laterally shifted by one of the eyes with respect to the other. As we will see later, the binocular disparity is inversely proportional to the distance of the object from the viewer.

With such a visual system aiding the perception of distance, the prey had a better chance of spotting an encroaching predator and escaping, and the predator had a better chance of triangulating the position of a prey and hunting. Binocular vision is thus believed to be a key enabler of evolutionary successes, and an attribute of the earliest mammals. In the modern era, we use our sophisticated binocular visual system to interact in the 3D world. Figure 1.3 also shows a simplified depiction of the sensory pathways connecting the eyes to the visual cortex.

Next, let us consider the key aspects of auditory perception, including the ears and the hearing process. Like the eye, the human ear has an elegant construction, with some astounding capabilities as the sound sensor. The ear, our natural *microphone*, is sensitive to over 12 orders of magnitude in sound intensity, and three orders of magnitude in sound frequency (20 Hz – 20 kHz)!

As depicted in Figure 1.4, the ear flap, also called the *pinna*, directs the air waves carrying the sound signals into the auditory canal that includes the eardrum or the *tympanic membrane*. The pressure vibrations are amplified through the middle ear components, the *malleus*, *incus* and *stapes*, which are the smallest bones in the human body and are also referred to as the *hammer*, *anvil*, and *stirrup*, indicative of how they amplify and transmit the sound signals

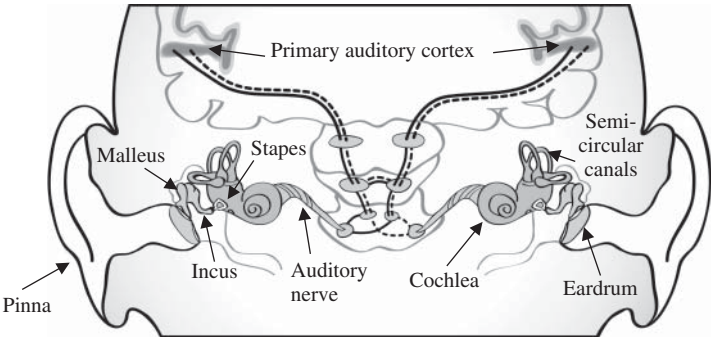

Figure 1.4 The anatomy of the human ear and the binaural construction, illustrating the simplified neural pathways between the cochlea in the inner ears and the auditory cortex in the temporal lobe of the brain. The binaural signals, along with frequency cues, are used for localization of the source for the sound signals.

to the inner ear section. Finally, the vibration waves of the sound are converted into neural signals via nerve impulses in the inner ear, more specifically by the hair cells at the converging spiral-shaped *cochlea*. These neural signals are subsequently dispatched to the auditory cortex of the brain, located in the temporal lobe, and are processed for perception.

As in the case of eyes, we also have two of these natural *microphones*, enabling a binaural perception scheme that, in addition to frequency cues, is capable of localizing the sources of sound accurately within the 3D space. While the binaural 3D perception, along with the extreme pressure sensitivity, was crucial to our evolutionary success, it is also crucial in helping us navigate and interact with the 3D physical world in our daily lives today. Figure 1.4 also shows the simplified neural pathway between the ears and the auditory cortex in the brain.

Finally, we turn to touch sensitivity and the tactile perception process. The perceptual process for our sensation to touch, also called the cutaneous sense, starts with the *mechanoreceptors* within the skin that pick up mechanical pressure in the corresponding skin area due to contact and activate neural responses. Figure 1.5 depicts the four principal types of mechanoreceptors. While the sensory organs for vision (the eyes) and for hearing (the ears) are located in the skull, with relatively short neurophysiological pathways to the cortex, the sensory organ for touch (the skin) covers the entire body. As a result, the signals from the touch receptors often have to travel long distances (e.g., from the fingertips to the head). The spinal cord serves as the "information highway" for the touch sensors, relaying the signals from the receptors to the *somatosensory* cortex in the parietal lobe, the part of the brain in the top region of the head that processes sensation to touch.

The seminal work of the neurosurgeon Wilder Penfield in the 1950s on tactile sensitivities has shown that adjacent parts of the human body map to adjacent areas of the cortex [11]. More interestingly, this mapping study has established the relative proportions of the somatosensory cortex that are dedicated to various body parts. This is presented by the concept of the *cortical homunculus*, as shown in Figure 1.5. Not to be mistaken as an arbitrary caricature, the homunculus depicts a scale model of the human body that represents the relative spaces that the corresponding body parts occupy on the somatosensory cortex. As the figure shows, the cortex dedicated to processing touch signals from the fingers far outweighs that dedicated to the entire arm and wrist, a vindication for the touch-screen based user-interface designers who

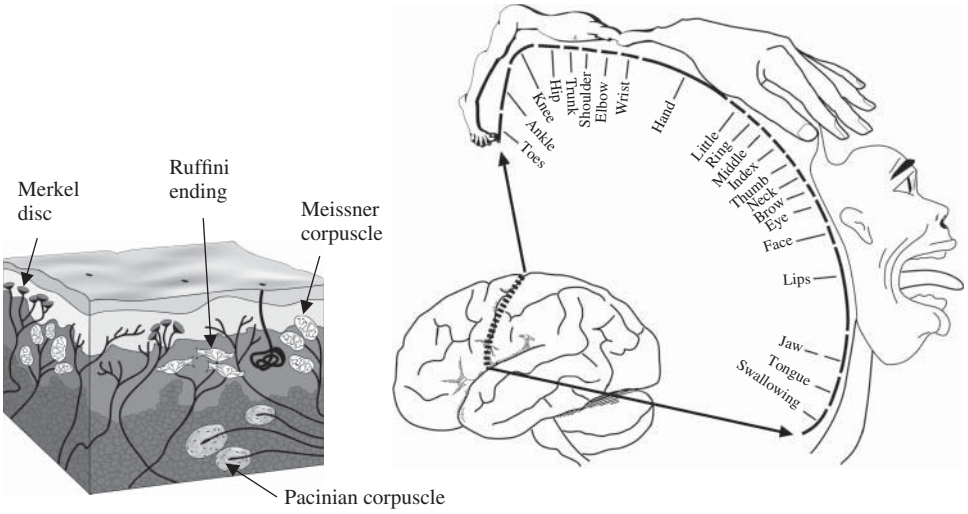

Figure 1.5 Left, anatomy of the human skin. The four principal types of mechanoreceptors that are shown convert the mechanical pressure of the tactile stimulus into neural signals. Right, the cortical homunculus, originally depicted by Wilder Penfield and refined by subsequent researchers, indicating the location and the relative portions of the somatosensory cortex dedicated to processing tactile signals from each part of the body. Source: http://www.intropsych.com/ch02_human_nervous_system/homunculus.html, reproduced with permission from R. Dewey.

assume an abundant use of the fingers for human interactions on touch-enabled interactive displays and devices!

So, as the preceding discussion reveals, there is a common theme to the neurophysiology of our perceptual processes. Our sensory systems are cleverly designed, with disproportionately large parts of the cerebral cortex assigned to the most important parts of the perceptual sensors – for example, the fovea of the eyes for central vision, hair cells in the cochlea of the ears for hearing, and the tips of the fingers for touch interaction. While we also possess other sensation mechanisms, we predominantly use vision, audition, and touch for interacting with the physical world surrounding us. Thus it is appropriate that in this book we focus on eyes, ears, and touch as the primary modalities of natural human interfaces to the interactive displays and devices.

In contrast to the biological systems, most of the computing and entertainment devices of today have rather rudimentary perceptual sensing and processing capabilities. For example, let us consider mobile phones, tablets, and laptop computers. They are typically "one-eyed" (with a single camera), like the Cyclopes of Greek mythology. In addition, most of them are monaural (with a single microphone), and many are yet to get tactile sensitivity (touch screen), especially among the laptop computers.

However, this situation is expected to improve in the near future, with rapid technical advances on many fronts. Taking a page from nature's playbook, device architects and designers are now starting to consider adding "human-like" sensing and perception capabilities to computing and communications devices, to give them the abilities to "see", "hear", and "understand" human actions and instructions, and use these new capabilities to enable natural

and intuitive human interactions. These developments promise to advance human-computer interactions beyond the confines of keyboard, mouse, joysticks, and remote controls, and allow the use of natural interactions built on touch, vision, and speech sensing and recognition technologies.

It is only when we attempt to implement such perceptual capabilities in machines that we realize how complex the task of sensation and perception is, despite the fact that we sense and perceive the world around us with surprising ease and casualness, at every moment of our conscious lives! In the next section, we review the key technologies for human interfaces with electronic devices – both the legacy technologies that have been widely adopted over the past decades, and also the recently emerging modalities of interactions with the displays and systems in natural and intuitive ways.

1.3 Human Interface Technologies

1.3.1 Legacy Input Devices

Before delving into the emerging natural interface technologies and the new class of applications and user experiences enabled by them, it behooves us to take a look at the history and reflect on the most successful user input technologies that have formed the backbone of human-machine interactions as we have come to know and practice them in our daily lives over the last few decades. A complete account of all human interface technologies and associated historical developments is not our intent, as it will be impossible to undertake such an endeavor within the practical limits of this chapter. A number of comprehensive reviews have been published in the past, and we point interested readers to them [13–15].

Here, we briefly cover a few of the seminal inventions and mainstream commercial implementations that have found adoption by the masses, and have defined human interactions with displays in major ways leading up to the modern era. As we look back, we appreciate the key reasons for their successes – simplicity in the technical implementations, utilizing the best of available technology ingredients within affordable price points and, above all, fulfilling user needs that helped enrich human lives and activities at the time of the particular invention and beyond.

First, the ubiquitous remote control device which has arguably defined our relationship and interactions with the television displays and shaped our content viewing behavior on it. While the concept of a remote control was described by Nikola Tesla as far back as 1898 [16], the first television display remote control was developed and commercialized by the Zenith Radio Corporation in 1950, who aptly named the device the "*Lazy Bones*" [17].

Televisions had been commercially available since the 1920s but they required people to walk up to them to manipulate the control knobs, while the natural content viewing position would have been sitting on a sofa in front of it. So, in retrospect, the environment was ripe for the invention of the remote control; the need was clear and, as we will see, technology ingredients were available. Zenith's "*Lazy Bones*" hand-held controller connected to the television with a long cable. While this solved a legitimate human need and allowed changing of television channels without having to get off the couch, the inadvertent tripping over the cable also pointed to the need to go wireless.

In came the "*Flash-matic*" in 1955, also from Zenith, which used a beam of light pointed to the sensors located in the four corners of the television set to control it remotely and

wirelessly. The excitement on it can be gleaned from a magazine advertisement from the day that boasted, "*You have to see it to believe it!*" Despite the enthusiasm, the light-control mechanism did not perform well in bright rooms, with the ambient light occasionally changing the settings. Zenith addressed the problem by switching to using ultrasound as the means of remote communication in their next device, named the "*Space Command*". An advertisement from 1957, shown in Figure 1.6, eloquently described, "*New miracle way to tune TV from your easy chair by silent sound.*"

Modern remote controls have come a long way since then, with fashionable and sleek form factors, utilizing infrared light to control entertainment devices. In recent years they have increasingly featured motion-sensing and voice control capabilities.

Next, we turn to the invention of the computer mouse by Douglas Engelbart, in 1963, which marked the beginning of a new era of human-computer interactions. Prior to the invention and deployment of the mouse, inputs to the early computers were limited to text-based commands typed on a keyboard. Figure 1.7 shows the first mouse prototype built by Engelbart and Bill English. It consisted of two wheels which rotated in mutually perpendicular directions as the mouse was dragged on a surface, allowing relative tracking of the location of the mouse on a 2D plane [18]. The idea of such a device construction occurred to Engelbart in 1961 as he was sitting in a computer graphics conference, pondering how to make a system that would allow easy and efficient interactions with the graphical objects on the computer screen [19].

It is noteworthy that the mouse was just of one of many mechanisms for computer input that Engelbart and his team at the Stanford Research Institute experimented with, albeit the most successful one in retrospect. Even though modern versions have replaced the wheels with a

Figure 1.6 An advertisement for the Zenith "*Space Command*" remote control device from 1957. Source: www.tvhistory.tv, reproduced with permission.

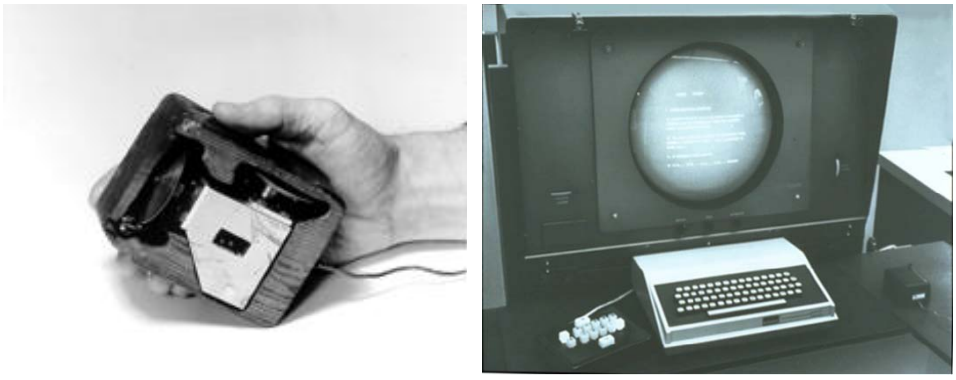

Figure 1.7 Left, the first computer mouse constructed by Douglas Engelbart and Bill English in 1963. Right, a computer workstation with an early mouse connected to it for user interactions Source: SRI International, reproduced with permission.

laser beam and replaced the cord with wireless transmission, we still refer to the device as a "mouse", which originally got its name from the resemblance to the namesake mammal due to the tail at the back to connect with the computer. The mouse, in conjunction with graphical user interfaces, made operation of computing systems easy for the user, and thus became ubiquitous along with the rapid proliferation of the personal computer over recent decades.

It is hard to overstate the scale of the impact of these "legacy" human interface devices on the proliferation of their respective host systems. As the television became the centerpiece of entertainment in the homes around the world, and the personal computer became the primary vehicle for productivity and information, the remote control and the mouse became our unavoidable companions. However, despite their widespread use over decades, the human interfaces and interactions with these devices are very limited, and it is time to look forward. As we will see, recent advances spanning new sensor technologies, inference algorithms, computing resources, and system integration have brought us to a point where we can start to realize natural human interfaces for interacting with electronic devices and systems. We now turn our discussion to user inputs with natural interfaces, based on touch, voice, vision, and multimodal interaction technologies.

1.3.2 Touch-based Interactions

The recent evolution of the display from merely presenting visual information to the user into an interactive interface device is largely due to the integration of touch-sensing capability into the display module, especially on the mobile devices. It took decades of development from the first report of capacitive touch screens, by Johnson in 1965 [20, 21], for the technology and usages to go mainstream with consumers worldwide.

The "touch display" constructed by Johnson, a cathode-ray-tube monitor with a capacitive touch overlay, is shown in Figure 1.8. Johnson, an engineer at the Royal Radar Establishment in England, developed the technology for use in air traffic control systems. The abstract of Johnson's paper stated, *"This device, the 'touch display', provides a very efficient coupling*

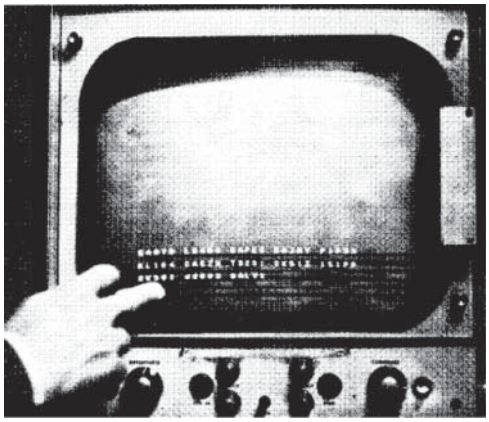

Figure 1.8 Image of the first "touch display" using the capacitive technique for sensing the contact location, reported by E. A. Johnson in 1965. Source: www.history-computer.com, reproduced with permission.

between man and machine." While Johnson's statement was specific to the device he constructed, it proved to be a profound foretelling of the devices and applications to come decades later, which would be adopted by the masses.

While the mouse and remote control devices enjoyed volume adoption for many years, interacting with the content on the display using these devices is necessarily an "indirect" manipulation experience. On the other hand, the introduction of touch-sensitive displays allowed people to "directly" interact with objects on the display by simply touching them, leading to easier and more intuitive experiences that required little or no training. In recent years, touch-screen technologies and their commercial deployment have been going through a tremendous phase of adoption and growth, thanks to the widespread proliferation of smartphones, tablets, ultrabook laptops, all-in-one desktop computers, and information kiosks of all form factors. In fact, the user experiences on these devices and systems have undergone a radical change due to the seamless integration of touch-screen technology and touch-friendly software interfaces, which have brought a new class of highly interactive applications to the users.

There are many different technical approaches to detecting touch inputs on a display device [22]. In Chapter 2, Walker provides an in-depth review of various touch technologies used in interactive displays. As described in detail in the chapter, methods for sensing contact location on the surface of a display can be predominantly categorized into capacitive, resistive, acoustic, or optical techniques.

In the capacitive approach, a layer is placed on or within the image display that stores electrical charges. In one of the implementations, called the mutual-capacitance method, when the user touches any location on the display, some part of the charge is transferred out into the user, resulting in a decrease in the electrical charge originally stored in that location. In another approach called the self-capacitance method, contact by a human body part increases the capacitance of a single electrode with respect to the ground. Electronic circuits designed to detect these changes within the touch-sensing layer are used to identify the location of the touch and provide this information to the software operating systems, applications and user interfaces.

In the resistive touch-sensing approach, when a user applies a mechanical force by touching a location on the display, two optically-transparent conductive surfaces, originally separated by a small space between them, come into contact at the touch location. The coordinates of this location are determined by voltage measurements and provided to the software for processing.

Acoustic and optical techniques involve measurements of the changes in ultrasonic waves and infrared optical waves, respectively, due to the touching of the surface of the display by the user. The specific system implementations widely vary across a wide range of products developed by a number of companies around the world.

The specific technologies covered in the chapter include projected capacitive, analog resistive, surface capacitive, surface acoustic wave, infrared, camera-based optical, in-cell integration, bending wave, force-sensing, planar scatter detection, vision-based, electromagnetic resonance, and combinations of these technologies. Walker presents the working principles of the various technologies, the associated system architectures and integration approaches, major advantages and disadvantages for each method, as well as the historical accounts, industry dynamics, and future outlook for the aforementioned touch-sensing technologies, including a discussion on various levels of integration of touch functionalities within the display. While many of the devices that are commercially available today use a touch-sensitive screen on top of the display module, recent developments and commercial products have demonstrated that touch-integrated displays without requiring separate touch-screen module to be attached on the panel provide a path to reducing the thickness, weight, integration complexity, and cost. Walker's chapter also covers such "embedded touch" technologies in detail.

The introduction of touch inputs to interactive displays and systems had a profound impact in the market. Let us take a quick look at the market size to gain appreciation for the footprint of touch-screen technology. The industry ships well over a billion units of touch-screens every year. While most of these are with mobile devices, touch screens are widely being adopted in devices of all form factors. It is just a matter of time before most displays will have touch input capabilities, especially the ones that require close-range user interactions.

1.3.3 Voice-based Interactions

Arguably, the most effective and prevalent form of communication and interaction between human beings is based on spoken language. To appreciate this, just do a "thought experiment" where you are a maverick world explorer and suddenly find yourself in a place among people who do not understand a thing you say and vice versa! Communicating with voice and speech has been a fundamental enabler of the modern human civilization and social interactions. Justifiably, there has been significant interest and efforts in the academic research community as well as the industry on developing human-computer interfaces utilizing voice input, processing, and output [23].

While we speak and understand what others are saying (mostly) without effort, making a computer understand spoken language by humans is a nontrivial task, and the effort has spanned a century. A quick look at Figure 1.9 gives us a glimpse of just a small part of the challenge, which shows a typical speech waveform recorded for the utterance of the phrase "*mining a year of speech*". When we speak sentences, we use non-uniform segmentations or temporal spacings between utterances and often generate an acoustic signal with no breaks where we perceive one, or with breaks in places that we do not perceive! We also often speak incomplete sentences in conversations, and mix meaningless utterances as "bridges" between

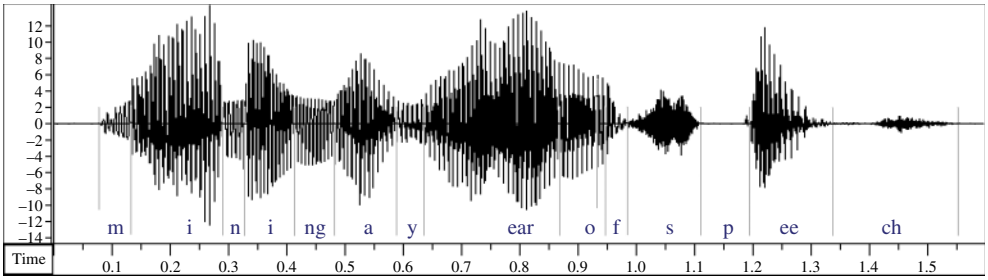

Figure 1.9 Typical speech waveform for the phrase "mining a year of speech", indicating the non-uniform temporal spacing between utterances and the nonintuitive breaks in the acoustic signal. The horizontal axis shows time in seconds and the vertical axis shows the signal strength in arbitrary units. Source: Reproduced from http://www.phon.ox.ac.uk/mining_speech/, with permission from John Coleman.

sentence segments. Broadly, the task of a speech recognition algorithm is to translate the spoken utterances into a series of text and, subsequently, to extract the meaning conveyed by them. A speech-based user interface uses these capabilities to establish a voice interaction scheme between the user and the device.

Since the early pioneering work in the 1920s and 1930s by Harvey Fletcher and Homer Dudley in Bell Laboratories on human speech modeling and synthesis [24, 25], the research on automatic speech recognition advanced steadily over decades, especially through the introduction of statistical algorithmic approaches in speech modeling in the 1980s, and the more recent advances in natural language understanding. In the 1968 epic science fiction movie *2001: A Space Odyssey*, screenwriters Stanley Kubrick and Arthur C. Clarke envisioned the HAL 9000, a computer that would be born in the 1990s and communicate with humans with fluent and natural conversational speech. While we are yet to achieve the marvelous feats that HAL was imagined to be capable of, recent developments in the domain of voice interfaces and interactions are starting to yield commercial success, with an increasing number of applications in mobile devices, consoles, and automotive usages.

In many instances, voice-based interfaces to computing devices result in simple and intuitive interactions between the humans and the devices. For example, a simple command in spoken language such as "*play the song [song title]*" would instantly pick out that specific song from among many songs stored in the device or a server and start to play it back. Similarly, a command such as "*post this picture to my Facebook page*" would instantly upload the picture that the user just captured using his or her smartphone, or one that he or she selected from the images stored in the device. "*Play the Wimbledon match saved from last night*" would find the tennis match from the media storage device and start playing it on the television screen. "*Give me the directions to SFO*" would display the map and driving directions to the San Francisco International Airport.

Accomplishing these tasks with traditional interfaces would require flipping through multiple command windows, typing in text inputs, and many clicks or taps. On the other hand, the same tasks completed via speech commands would be inherently easier and faster, if the devices are capable of understanding and interpreting voice commands and instructions with the required accuracy levels in the real-world usage conditions.

Voice recognition with natural language understanding and speech synthesis promise to expand significantly the usages for computing, communications, entertainment, and a

plethora of other electronics devices and systems. It makes the use of devices possible when the hands and eyes are occupied in other tasks such as cooking, driving, shopping, gardening, jogging, etc., and potentially makes computing tasks feasible for many disabled individuals.

As we will discuss later, voice-based interactions are especially powerful when used along with other interface methods such as gesture or eye gaze tracking. In the future world of truly ubiquitous computing, when sensing and inference technologies become embedded into everything around us – from the things we wear to the things we carry to the places we live and work in – interactions based on spoken language will be crucially important. For now, from the viewpoint of interactive displays, voice interfaces seem poised to make the experiences of interacting with the content on displays of all form factors easier and more intuitive.

In Chapter 3, Breen *et al*. present an in-depth review of the fundamentals and developments in voice-based user interfaces. The authors provide a thorough discussion on the key elements of speech interfaces, including speech recognition, natural and conversational language understanding technologies, dialog management, speech synthesis, hardware and system architecture optimization for efficient speech processing, and applications to a broad array of interactive devices and systems.

1.3.4 Vision-based Interactions

As we discussed in Section 1.2, visual perception, more specifically the ability to see and understand the 3D environment, is a crucially important capability that enables us to navigate with ease in the physical world around us, and to interact with it and each other. 2D cameras and imaging applications are already ubiquitous part of computing and entertainment devices now – especially mobile phones, tablets, and laptop computers – and are increasingly appearing in all-in-one desktop computers and even in the high-end large-screen televisions.

Currently, the primary application for 2D cameras integrated within mobile devices is the capturing of digital still photographs and videos, while those in larger devices and displays are mostly used for video conferencing applications. Computer vision researchers have also developed 2D image processing algorithms that can detect, track, and recognize faces and facial expressions, understand poses and simple gestures [26–29].

A traditional 2D camera captures the projection of the 3D world onto the 2D image plane, and discards a wealth of other visual information in the 3D space in front of it. There have been significant research efforts on recovering the 3D information from single 2D images to understand human poses. Reconstructing 3D spatial information from 2D projection is an ill-posed problem with inherent ambiguities, and is a challenge even for fitting of known skeletal structures such as the human body, even though promising results have been reported for limited uses [30–32]. These approaches are in general computer intensive and often require manual user inputs, and hence are not suitable for applications that require real-time and unaided understanding of the 3D environment and general human gestural interactions in the 3D space.

In contrast, the 3D imaging pipeline in the human visual system captures and utilizes 3D visual information to enable efficient and robust cognition and interaction. Adding real-time 3D visual perception capabilities can enable truly interactive displays and systems that are capable of understanding the users and the rich natural user interactions. These include real-time 3D image sensing techniques to capture the 3D scene in front of the display; computer vision algorithms to understand the 3D images and real-time user activities in the

3D space; and appropriate user interfaces that couple human actions and instructions to the system intelligence and responses in intuitive ways.

Vision-based gesture recognition is a burgeoning field of research and development across the world, with rapidly advancing techniques reported by both academic and industrial labs, in conjunction with developments in classifying and implementing various levels of interactions based on the study of human motion behaviors [29, 33–35]. Chapter 4 provides an overview of vision-based interaction methods, including 3D sensing and gesture recognition techniques, status and outlook for implementing them in human-computer interaction applications.

Systems and applications based on 3D sensing devices have started to appear in the market, which offer richer and more robust interactive experiences than those implemented on the traditional 2D imaging methods [36, 37]. These early commercial successes promise to propel the adoption of 3D vision technologies into a wide array of devices and systems in the future, thereby making 3D user interactions pervasive. Adding real-time 3D imaging technology to electronic devices enables fine-grain user interactions and object manipulations in the 3D space in front of the display.

There are various methods for real-time 3D sensing, all of which generally output a depth-map for the scene in addition to a color image, allowing reconstruction of the 3D objects and scenes that were imaged. Three of the most prominent methods are: structured-light 3D sensing techniques, stereo-3D imaging, and time-of-flight range imaging techniques [37]. Chapters 5–7 delve into the details of each of these specific 3D imaging techniques that provide the foundation for implementing 3D interactive applications.

With real-time acquisition of 3D visual information using the techniques described above, rich human-computer interaction schemes can be implemented using 3D image recognition and inference techniques that enable interactions beyond touch screens. Figure 1.10 shows some examples that are naturally enabled by 3D gesture interactions in front of the display, rather than the use of traditional 2D inputs such as mouse or touch-screen [37]. The image on the left shows a scenario where the user is expected to reach out and "grab" a door knob, "turn" it, and "pull" it out of the plane of the display to "open" the door. The image on the right shows a "slingshot" application, where the user "pulls" a sling with the fingers, "directs" it in the 3D space, and "releases" it to hit and break the targeted elements of a 3D structure. These actions would clearly be quite difficult to implement with a mouse, keyboard, or a touch-screen, and would not be intuitive experiences for the user. However, implementations of 3D gesture interactions using real-time 3D image capture and 3D computer vision algorithms result in more natural and intuitive user experiences for this type of applications.

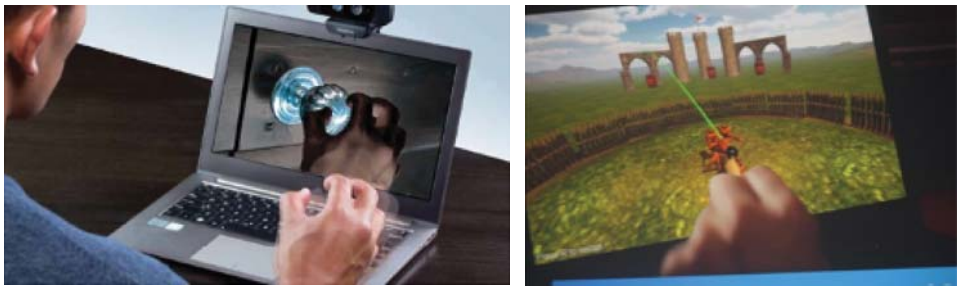

Figure 1.10 Examples of interactive applications and experiences enabled by real-time 3D sensing and inference technologies include manipulations of objects in the 3D space in front of the display [37].

Besides gesture interactions and object manipulations in the 3D space, real-time 3D imaging can also transform photography, video conferencing, remote collaboration, and video blogging applications. For example, the users can be segmented in the images more easily and accurately using the depth map generated by the 3D imaging devices, and subtracted from the background or placed in front of a custom background. This is illustrated in Figure 1.11.

While image processing techniques can be used to achieve this on traditional 2D images, 3D sensing devices allow cleaner segmentation and real-time applications utilizing the 3D scene information. For example, one could participate in a business meeting via a video conferencing application from the comfort of his or her home, but is shown on the screen in front of his or her office background instead!

Another category of applications that can be significantly enhanced is augmented reality, where rendered 3D graphical content is added to the captured image sequences. Beyond the traditional augmented reality applications that currently use 2D cameras, 3D imaging can augment video content with 3D models of objects and scenes for realistic visual representations, and allow the users to interact with the elements within the augmented world. Imagine applications that allow you to virtually try on clothes or jewelry in front of an interactive display equipped with a 3D imaging device, or virtually decorate your room to select appropriate furniture.

Beyond the tracking and recognition of hand and body gestures, there have also been significant developments in technologies that can detect eye gaze direction and determine where on the display the user is looking. Eye gaze plays an important role in inter-human interactions in our daily lives. Gaze is an important indicator of attention. For example, Figure 1.12 shows the regions of interest for a specific person while viewing an image.

Neurophysiological studies have shown the importance of gazes in consistent interaction with the physical world [38, 39]. While the primary function of the eye is to capture the visual information from the scene as part of the visual perception process, we also deploy eye gazes in close coordination with speech and gestures as we communicate and interact. As an example, as you say "*please give me that red ball*" and look at the red ball on the chair, it becomes

Figure 1.11 3D segmentation using a depth-sensing imaging device allows easy manipulation of the background. In this illustration, the boy is shown on the left in front of the original background, and on the right on a different background after post-processing. Note that an analysis of the inconsistent shading in the right image would reveal that the background is not original. Depth-sensing imaging devices enable real-time segmentation using the 3D scene information, for use in applications such as video conferencing or video blogging with a custom background. Source: Reproduced with permission from Sean McHugh. www.cambridgeincolour.com. (See color figure in color plate section).

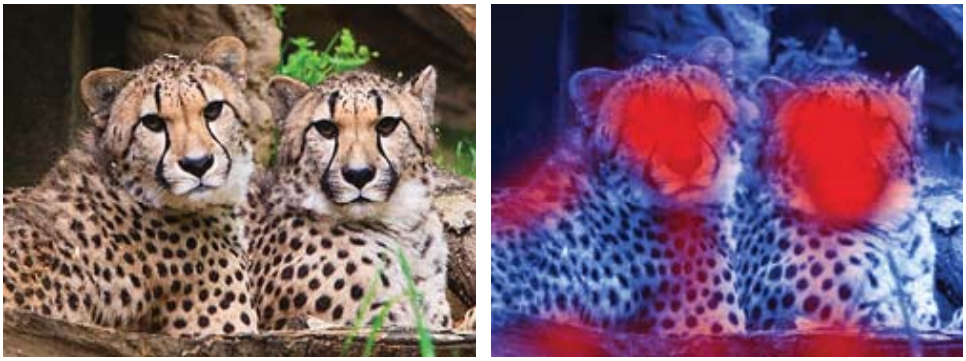

Figure 1.12 An example of visual attention indicated by the directions of eye gazes. Left, an image shown to the viewer. Right, regions of interest within the image. Source: cambridgeincolour.com, reproduced with permission from Sean McHugh. (See color figure in color plate section).

obvious to the person looking at you that it is not the red ball lying on the floor that you are referring to. The person can comprehend this by simply following the direction of your eye gaze, even if you are not using your finger to point to the ball on the chair.

Researchers have long been fascinated by the prospect of incorporating this powerful interaction mechanism into user interfaces with computing systems, especially in conjunction with other interaction modalities. For example, one could simply look at an icon on the screen of the laptop computer and say "open it" or "launch it", rather than reaching out to tap on the touch screen or using the mouse to point and click, or even performing a hand gesture in free space. In Chapter 8, Drewes provides an in-depth review of eye gaze tracking technologies, systems, and applications, including the current limitations of deploying gaze tracking in human-computer interaction schemes and potential paths to mitigate these challenges.

1.3.5 Multimodal Interactions

Human perception and interactions are often multimodal – we use all our senses and combine the neural signals generated by them in order to understand the physical world and interact with it. For example, we use binaural audio signals along with frequency cues to locate the source of a sound, and then use the convergence and accommodation mechanisms in the eyes to point our binocular vision towards it, and bring the light rays emanated by the object to focus on our retina in order to see it at the same time that we hear its sound. Similarly, on other occasions, auditory perception may follow visual perception and add to it. For example, while strolling in a park, we may first see a bird, then pay attention to the sound of its chirping. In the real world, we use multiple modalities of interactions to communicate with each other. Based on intent and context, we use combinations of touch, gesture, voice, eye gaze, facial expressions and emotions to interact intuitively with fellow human beings.

In a seminal paper in 1976, aptly titled *Hearing Lips and Seeing Voices*, McGurk and MacDonald narrated their serendipitous discovery of an interaction between vision and hearing that has since come to be known as the "McGurk effect" [40]. This work demonstrated that when the auditory element of a sound uttered by a person is accompanied by the visual element of a different sound with a dubbing process, it leads to the perception of a third

sound. The fusing of seeing and hearing in our perceptual processes is also evident in the ventriloquism effect, as well as in theaters where we get the illusion that the actors on the screen are speaking, even though the speaker systems are located elsewhere in the room. Neurophysiological evidence has established that the neural signals from one perceptual sensor can enhance, override, or modify those from the other as we comprehend our surroundings with multisensory perception. The interplay of different sensory areas within the human brain has been demonstrated, providing experimental evidence of connections between the receiving areas in the brain for vision, hearing, and touch [41].

Natural and intuitive human-computer interaction schemes must therefore be multimodal. The early results of combining speech recognition and position sensing, described by Bolt in his 1980 paper, demonstrated the feasibility of natural discourses between the human and the machine, such as "put that there", "make that a large blue diamond", "call that … the calendar", etc. [42]. Quek writes, *"For human-computer interaction to approach the level of transparency of interhuman discourse, we need to understand the phenomenology of conversational interaction and the kinds of extractable features that can aid its comprehension,"* and demonstrates the use of speech and gesticulation as coexpressive forms of communication [34].

In Chapter 9, LaViola *et al*. review multimodal perceptual interfaces in human-computer interaction, exploring the combination of various input modalities to constitute natural communications. The chapter examines the dominant interaction types, usability issues of various levels of multimodal integration, and methods to mitigate them towards achieving life-like natural interactions. Addressing human-factors aspects of multimodal interface schemes is crucial to the commercial success of new devices and systems incorporating multimodal interactive functionalities. Besides the input modalities covered in the preceding chapters, such as touch, gesture, speech, eye gaze, and facial expressions, this chapter also provides a discussion on integrating the emerging brain-computer interface technologies based on electroencephalography and muscle activity detection based on electromyography.

Science fiction authors have long fantasized about a future world where one controls computers, machines, and systems with brain waves, such that one simply thinks and things happen! While that future practically still remains to be realized, recent developments in brain interface technologies have demonstrated the ability to control and manipulate content on the display by signals emitted by the brain as a result of thoughts. Research and development continues in this domain, which promise to bring yet-to-be imagined interaction schemes and applications that will enrich future interactive displays and systems [43]. In Chapter 9, LaViola *et al*. consider integration of such brain-computer interfaces within the multimodal interaction schemes.

Besides multimodal interactions with the content on a display, the advances in face- and voice-based user recognition methods promise to replace the use of passwords for user identification by natural multimodal biometric authentication technologies. In our social interactions in daily lives, we use face, voice, and behavioral traits based human recognition schemes to establish the identity of people that we interact with. In contrast, a computer's ability to identify its users is still largely limited to passwords or tokens. As computing becomes pervasive and integrated in all aspects of our lives and the society, this is no longer going to be sufficient.

In Chapter 10, Poh *et al*. reviews multimodal biometrics, including technological design and usability issues, and recent developments in the field. As another example of multimodal perception, as we communicate with each other, we often use the cues present in facial expressions to understand verbal discourse. The same words, uttered with different expressions on the face, can mean very different things. Facial expressions can be voluntary to reinforce

communications via specific gestures sported on the face, or involuntary indications of the inner feelings and emotions. The interpretations of one's facial expressions by other observers are dependent on context [44].

More than 150 years ago, Duchenne conducted experiments on human subjects to study how the movements of muscles produce various expressions on the face. As an example from his work, Figure 1.13 shows a number of facial expressions that were imparted on the faces via muscle contractions induced with electrical probes, and recorded using the newly available camera device [45]. The advent of digital cameras, advanced image processing techniques, and computational resources over the last decades have made more natural studies of facial expressions possible. More recently, 3D sensing and processing techniques are increasingly being used for more advanced and automated recognition of facial expressions. A brief review of the developments in vision-based facial expression recognition techniques is included in the discussion on visual sensing and gesture interactions in Chapter 4.

1.4 Towards "True" 3D Interactive Displays

Although visual displays have become ubiquitous and indispensable parts of our lives, the vast majority of them currently serve the primary role of displaying monocular (2D) visual information and are unable to reconstruct important visual cues that are salient to our 3D perception of the real world. In the recent years, stereo-3D displays have also started to gain traction in the market. The major focus for the stereo-3D displays that are commercially available currently has been on providing the stereopsis cue, where different images are presented to the left and right eyes of the viewer to create depth perception by using the binocular fusion process in our visual system. There are a number of books that cover the principles of various display technologies to reconstruct both 2D and 3D imagery [1–5].

Figure 1.13 Original work of Duchenne, published in 1862, on inducing various facial expressions by muscular contractions activated by the application of electrical probes. The top figure shows the same expressions induced on both sides of the face, whereas the bottom figure shows different expressions induced on either sides of the face. Adapted from Duchenne [45]. Source: Reproduced with permission from www.zspace.com.

The ultimate goal is to construct "true" 3D interactive displays and systems that would provide life-like and immersive visual and interaction experiences to the user. The development of such displays requires careful examination of the human visual perception system and processes to reconstruct signals that are consistent with the visual cues that we utilize for sensing the 3D world in our day-to-day lives. So, how do we see and perceive the third dimension with our visual sensing and processing system?

Our 3D perception of the real world utilizes several important 3D visual cues, besides stereopsis. These include the motion parallax effect, where nearer objects appear to move faster across the view relative to objects that are further when the viewer moves; the convergence effect, where the eyes rotate inward or outward to converge on an object that is located nearer or further; the accommodation effect, where the shape of the lens in the eye is changed to focus on an object; the occlusion effect, where nearer objects partially hide objects that are further away; the linear perspective effect, where parallel lines converge at a distant point on the horizon; the texture gradient effect, where uniformly spaced objects appear to be more densely packed with distance; shadows that are consistent with 3D location of the objects and the lighting of the environment; and other cues arising from prior knowledge, such as familiar sizes, atmospheric blurs, etc. A number of these important 3D visual cues that contribute to our 3D perception are shown and explained in Figure 1.14.

It had been demonstrated that implementing motion parallax capability in a display to show unique views such that the images projected on the retina consistently varies with the head and eye movements of the viewer, in addition to stereopsis, provides a more lifelike visual experience. An example of such an implementation is a display from zSpace, illustrated

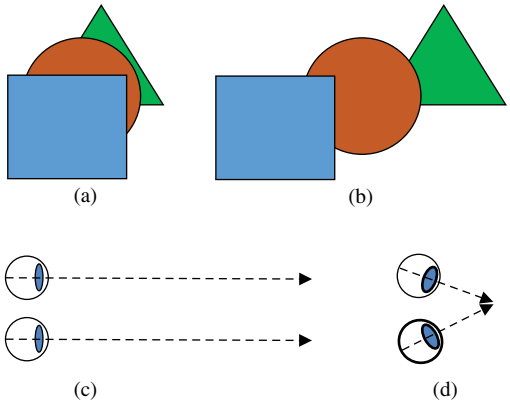

Figure 1.14 Depiction of some the salient visual cues that contribute to our 3D perception of the world, and which the future "true" 3D displays need to produce in order to provide immersive 3D visual experience. Top figures of the stacked shapes depict the binocular disparity with stereopsis cue: (a) is the image seen by the left eye of a viewer, and (b) is the image seen by the right eye. It also demonstrates the occlusion cue, as a single view is enough to indicate that the square shape is nearer to the viewer, whereas the triangle shape is further away. It also explains the motion parallax cue, since the square moves more across the visual field than the circle, which is further away, as the eye position moves from left to right. The bottom figures explain the convergence and accommodation cues: (c) depicts the case where the optical axes of the eyes are almost parallel to each other when seeing an object very far away; (d) eyes converge to see a nearer object, and the shape of the lens is accommodated to bring the image of the object to focus on the retina. Besides these, there are other 3D visual cues that are explained in the text.

Figure 1.15 Illustration of an interactive display made by zSpace that combines motion parallax effect along with stereopsis. The system tracks the head movement of the user and displays a stereo-pair of images that are unique to the viewer's position. A stylus is used for real-time interaction with the virtual objects on the display. Source: www.zspace.com, reproduced with permission.

in Figure 1.15. This 3D display system tracks the user's head movements with infrared camera sensors and renders a stereo-pair of images that are unique to the viewer's head position, thereby providing real-time motion parallax visual cue [46]. It also includes a stylus to manipulate the virtual objects in 3D space.

Traditional stereo-3D displays also suffer from inconsistent focus cues due to the mismatch between the convergence of the eyes to focus visual attention and the accommodation of the lenses to focus the incoming light. This conflict was shown to cause visual fatigue in human observers [47], and recently a way to mitigate this issue by using electrically tunable lenses has been proposed [48].

In the earlier sections of this chapter, we have discussed that the additions of touch sensors and associated user interfaces, especially on mobile displays, are turning displays into two-way interaction devices. We have also seen that beyond the touch inputs that are limited to the 2D plane of the displays, recent progress in 3D imaging and activity recognition techniques are increasingly allowing the implementations of user interactions in the 3D space in front of the display. We anticipate that the developments in these two fields will be combined to architect systems that will have end-to-end 3D user interfaces, simultaneously showing 3D visual content and understanding 3D user inputs.

Clearly, using a 2D user input scheme such as touch or a mouse to manipulate the visual content shown on a 3D display does not provide a natural or intuitive user experience, and using a 3D interaction scheme would be more appropriate. For example, studies of subjective user experiences have demonstrated significant issues with in-plane touch-based user interaction on the 3D visual content displayed by stereo-3D displays [49], whereas users demonstrated a tendency to interact with 3D virtual objects with gestures [50]. Although there is increasing interest in the research and development of direct 3D interaction schemes for manipulating objects shown on 3D displays [51–54], further developments are still ahead towards realizing practical mainstream implementations.

The future "true" 3D interactive displays will need to present dynamic 3D visual content that provides consistent stereopsis, parallax, and focus cues to the viewers for depth perception in addition to the monocular 3D cues, and at the same time include 3D sensing and inference technologies to allow immersive interactions with the objects reconstructed in the 3D space. Chapter 11 presents an in-depth analysis of the requirements and progress towards achieving this goal. This chapter first details the fundamental requirements to reconstruct "true" 3D visual information using the principles of light field and the basics of human visual perception process. It then reviews the progress in technology developments towards realizing such "true" 3D visual displays that would provide all the important visual cues that are necessary for lifelike 3D perception. Finally, the integration of human interactions with the 3D visual content on such displays and systems is addressed, including human factors issues and potential solutions.

1.5 Summary

Visual information displays are now everywhere. They are the face of all kinds of computing, communications, entertainment, and other electronic devices and systems. As a result of the relentless drive in technology developments over the past few decades to achieve high visual quality, a wide range of sizes with thin profiles, low power consumption and affordable price points, displays have evolved to deliver stunning visual performance. Commercial shipments have skyrocketed due to the rapid consumer adoption of devices of all form factors, ranging from wearable devices to handheld smartphones to tablets and laptops to large-screen televisions and information kiosks. Now, the displays are entering a new era, morphing from a one-way visual information device to a two-way interactive device.

Human-computer interaction schemes are going through a transformation as well, with the traditional keyboard and mouse interfaces being replaced or augmented by direct and natural input methods utilizing touch, voice, and gestures. Thanks to the rapid proliferation of touch-sensing technologies, the user experiences on mobile devices have literally been transformed, with exciting new user interfaces and applications. The recent advances in 3D computer vision with real-time 3D image capture techniques and inference algorithms promise to take it one step further, by enabling rich human-computer interactions in the 3D space in front of the display. In addition, significant technology advances have been demonstrated on speech interfaces, eye gaze detection, and brain-computer interfaces. Multimodal interactions, based on multisensory perception and combination of the various input modalities, promise to make the interaction experience lifelike.

This book presents an in-depth review of the technologies, applications, and trends in the rapidly emerging field of interactive displays, focusing on natural and immersive user interfaces. In this chapter, we have provided an overview of the sensing and perception processes that are relevant to the understanding and development of interactive displays, and reviewed the natural human interface technologies. Just as the introduction of the mouse and the graphical user interface a few decades ago brought about numerous new applications on the computers, and the proliferation of the touch interface enabled another set of new applications on the smartphones and tablets over the past few years, the natural and intuitive user interfaces based on 3D and multimodal sensing and inference technologies are all set to usher in a new class of exciting and interactive applications on computing, communications, and entertainment devices. The future of the displays is interactive, and that future is already here!

References

1. Bhowmik, A.K., Bos, P.J., Li, Z. (Eds.) (2008). *Mobile Displays: Technology & Applications*. John Wiley & Sons, Ltd.
2. Lee, J.H., Liu, D.N., Wu, S.T. (2008). *Introduction to Flat Panel Displays*. John Wiley & Sons, Ltd.
3. Brennesholtz, M.S., Stupp, E.H. (2008). *Projection Displays*. John Wiley & Sons, Ltd.
4. Tsujimura, T. (2012). *OLED Displays*. John Wiley & Sons, Ltd.
5. Lueder, E. (2012). *3D Displays*. John Wiley & Sons, Ltd.
6. IHS Displays Report Portfolio, www.ihs.com 2009–2013.
7. Goldstein, EB. (2013). *Sensation and Perception*. Cengage Learning.
8. Snowden, R., Thompson P., Troscianko T. (2006). *Basic Vision: An Introduction to Visual Perception*. Oxford University Press.
9. Lee, M., Jago, J., García-Bellido, D.C., Edgecombe, G.D., Gehling, J.G., Paterson, J.R. (2011). Modern optics in exceptionally preserved eyes of Early Cambrian arthropods from Australia. *Nature* **474**, 631–634.
10. Parker, A. (2011). On the origin of optics. *Optics & Laser Technology* **43**, 323–329.
11. Penfield, W., Rasmussen, T. (1950). *The Cerebral Cortex of Man: A Clinical Study of Localization of Function*. MacMillan.
12. Dewey, R. *Psychology: An Introduction*. www.intropsych.com.
13. Jacko, J.A. (Ed) (2012). *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. CRC Press.
14. Milner, N.P. (1988). A review of human performance and preferences with different input devices to computer systems. In: Jones DM, Winder R. (Eds). *Proceedings of the Fourth Conference of the British Computer Society on People and computers IV*, pp. 341–362. Cambridge University Press.
15. Buxton, W. *Human Input to Computer Systems: Theories, Techniques and Technology*. http://www.billbuxton.com/inputManuscript.html.
16. Tesla, N. (1898). *Method of and Apparatus for Controlling Mechanism of Moving Vessels or Vehicles*. US Patent 613,809.
17. A Brief History of the Remote Control. (1999). *DigiPoints: The Digital Knowledge Handbook* **3**, 4.
18. English, W.K., Engelbart, D.C., Berman, M.L. (1967). Display-Selection Techniques for Text Manipulation. *IEEE Transactions on Human Factors in Electronics* HFE-8, 5–15.
19. *Father of the Mouse*. http://dougengelbart.org/firsts/mouse.html.
20. Johnson, E.A. (1965). Touch Display – A novel input/output device for computers. *Electronics Letters* **1**, 219–220.
21. Johnson, E.A. (1967). Touch Displays: A Programmed Man-Machine Interface. *Ergonomics* **10**, 271–277.
22. Bhalla, M., Bhalla, A. (2010). Comparative study of various touchscreen technologies. *International Journal of Computer Applications* **6**(8), 975–8887.
23. Pieraccini, R., Rabiner L. (2012). *The Voice in the Machine: Building Computers That Understand Speech*. The MIT Press.
24. Fletcher, H. (1922). The Nature of Speech and its Interpretations. *Bell Systems Technology Journal* **1**, 129–144.
25. Dudley, H. (1939). The Vocoder. *Bell Labs Record* **17**, 122–126.
26. Yang, M., Kriegman, D., Ahuja, N. (2002). Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(1), 34–58.
27. Tolba, A, El-Baz, A, El-Harby, A. (2006). Face Recognition: A Literature Review. *International Journal of Signal Processing* **2**(2), 88–103.
28. Fasel, B., Luttin, J. (2003). Automatic Facial Expression Analysis: a survey. *Pattern Recognition* **36**(1), 259–275.
29. Mitra, S., Acharya, T. (2007). Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews* **37**, 311–324.
30. Lee, H., Chen, Z. (1985). Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing* **30**, 148–168.
31. Guan, P., Weiss, A., Bălan, A., Black, M. (2009). Estimating Human Shape and Pose from a Single Image. *Int. Conf. on Computer Vision* 1381–1388.
32. Ramakrishna, V., Kanade, T., Sheikh, Y. (2012). Reconstructing 3D human pose from 2D image landmarks. *Proceedings of the European Conference on Computer Vision, Part IV* 573–586.
33. Pavlovic, V., Sharma, R., Huang, T. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence* **19**(7), 677–695.

34. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., McCullough, K.E., Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction* **9**, 171–193.
35. Wexelblat, A. (1995). An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction* **2**, 179–200.
36. Han, J., Shao, L., Xu, D., Shotton, J. (2013). Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics* **43**(5), 1318–1334.
37. Bhowmik, A. (2013). Natural and Intuitive User Interfaces with Perceptual Computing Technologies. *Inf. Display* **29**, 6.
38. Pelphrey, K.A., Morris, J.P., McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain* **128**, 1038–1048.
39. Klin, A. Jones, W., Schultz, R., Volkmar F. (2003). The enactive mind, or from actions to cognition: Lessons from autism. *Philosophical Transactions of the Royal Society of London B* **358**, 345–360.
40. McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* **264**, 5588.
41. Murray, M., Spierer, L. (2011). Multisensory integration: What you see is where you hear. *Current Biology* **21**, R229–R231.
42. Bolt, R. (1980). Put-That-There: Voice and Gesture at the Graphics Interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques* 262–270.
43. Tan, D.S., Nijholt, A. (2010). *Brain-Computer Interfaces and Human-Computer Interaction*. Springer-Verlag.
44. Carroll, J., Russell, J. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology* **70**, 205–218.
45. Duchenne, G. (1862). The Mechanism of Human Physiognomy (Mecanisme de la physionomie Humaine).
46. Flynn, M., Tu, J. (2013). Stereoscopic Display System with Tracking and Integrated Motion Parallax. *International Display Workshops* **3**, D1–3.
47. Hoffman, D., Girshick, A., Akeley, K., Banks, M. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision* **8**(3), 1–30.
48. Bos, P.J., Bhowmik, A.K. (2011). Liquid-Crystal Technology Advances toward Future True 3-D Flat-Panel Displays. *Inf. Display* **27**, 6.
49. Pölönen, M., Järvenpää, T., Salmimaa, M. (2012). Interaction with an autostereoscopic touch screen: effect of occlusion on subjective experiences when pointing to targets in planes of different depths. *Journal of the Society for Information Display* **20**(8), 456–464.
50. Grossman, T., Wigdor, D., Balakrishnan, R. (2004). Multi-finger gestural interaction with 3d volumetric displays. *Proceedings of the 17th annual ACM symposium on User interface software and technology,* 61–70.
51. Alpaslan, Z., Sawchuk, E. (2004). Three-dimensional interaction with autostereoscopic displays. *Proceedings of SPIE Vol. 5291A, Stereoscopic Displays and Virtual Reality Systems XI* 227–236.
52. Blundell, B. (2011). *3D Displays and Spatial Interaction: Exploring the Science, Art, Evolution and Use of 3D Technologies*. Walker & Wood Ltd.
53. Bruder, G., Steinicke, F., Stuerzlinger, W. (2013). Effects of Visual Conflicts on 3D Selection Task Performance in Stereoscopic Display Environments. *Proceedings of IEEE Symposium on 3D User Interfaces 3DUI*. IEEE Press.
54. Zhang, J., Xu, X., Liu, J., Li, L., Wang, Q. (2013). Three-dimensional interaction and autostereoscopic display system using gesture recognition. *Journal of the Society for Information Display* **21**(5), 203–208.