Vision-Based HCI Applications

Eric Petajan

face2face animation, inc.
eric@f2f-inc.com

Vision-based HCI promises to simultaneously provide much more efficient communication from human to computer, and increased security using biometric identity verification. This chapter describes mature and currently deployed applications while offering reasons for slow deployment. System architecture and social issues are also explored resulting in the recommendation of a client-server architecture with standards like MPEG-4 Face and Body Animation (FBA) for optimal resource utilization; especially given the rapid adoption of powerful mobile devices as the primary HCI device.

1 Introduction

A growing number of people are spending more and more time interacting with electronic machines that are increasingly mobile, wireless, and compact. The world is in love with mobile devices as evidenced by the one billion phones in use today. While personal computer users will continue to spend time with larger fixed displays or medium sized portable displays, many times more people will own mobile phones with small displays. Eventually, the choice of display size will be independent of the choice of content. This independence is well on its way to realization as we watch video on our phones and drive our flat screen HDTV displays with computers. The cost of mobile web access is dropping rapidly while handsets are offered with a range of display sizes from tiny to a handful. The drop in cost of HD displays has caused a proliferation of flat panel displays in public spaces, the workplace, vehicles, and, of course, the home. In general, the flow of information from machine to human has progressed steadily with improvements in display technology, graphics chips, bandwidth, and battery life. However, the flow of information from human to machine is still mostly limited by the keyboard and mouse for the input of symbolic and spatial information. This one-way bottleneck in communication is especially constricted with mobile devices where keyboards are not very practical and require much more effort to use. If one's hands or eyes are busy while driving, walking, or handling something, the use of all but the simplest tactile interfaces is impractical or even hazardous for most people. Fortunately, the audio-visual computer input modalities can be used when the hands, eyes and ears are busy. Furthermore, the physical size of cameras, microphones, and processors will eventually be smaller than the smallest mobile display.

Each HCI modality has distinct advantages and limitations. An optimal HCI system should provide the user with the right combination of tactile, audio and visual modes given the amount of mobility and information exchange required at the time. The visual input modality is the least developed due to hardware cost and system complexity. The audio/speech input modality is still not reliable enough for widespread use, but integration with visual speech and gesture [1] recognition should significantly increase tolerance for audio mode recognition errors. Speech and gesture recognition are both the most widely used and natural human communication modes, and the least supported in current HCI systems. Most people can speak much faster than they can type and if Automatic Speech Recognition (ASR) was faster and more reliable it would be widely used for HCI. The incorporation of visual speech recognition into ASR promises [2, 3, 4] to provide sufficient robustness for general use. Simultaneously, face and voice recognition could also be deployed to identify the user. Finally, as vision-based HCI becomes a personal appliance that is always on and networked, audio and visual privacy will need to be secured and reliably controlled by the user. As the other electronic components continue miniaturization, the size and power consumption of the display will eventually be the first consideration when the user chooses an information appliance. Power consumption can be minimized at any time by placing as many applications on servers as possible while treating the personal appliance as a thin client.

While this book deals with vision-based interfaces, the audio modality is a necessary component of the ultimate HCI. Since the human face is the center of human communication, a primary requirement of vision-based HCI systems must be the capture and understanding of speech and emotional state using both audio and visual modalities. These two modes are uniquely suited for the acquisition of human behavior in that no physical contact is required, freeing the hands to perform other tasks. However, the lack of physical contact between human and device increases the possibility that audio or visual information about or associated with the user will be inadvertently transmitted to unintended recipients. Given the fallible nature of human beings and the lack of security inherent in legacy networks and computers, a secure privacy solution will necessarily be implemented as part of the local audio/visual acquisition system. It should also provide the user with continuous feedback and training for optimal positioning, voice level, and use of gestures, while maintaining awareness of the user's presence and identity. Limitations, some temporary and some fundamental, have impeded the realization of this utopian HCI system. This chapter explores how the receding of limitations should result in significant progress toward vision-based HCI for the masses, while certain applications are enabled in the short term.

2 System Architecture Considerations

The simultaneous demand for mobility, access to resources and information, and visual privacy points to the use of a client-server architecture where the HCI is part of a thin client with reliable data communication to the server. Given that state-of-the-art video codecs can't compress video enough for low latency transmission over consumer networks, the vision processing must be performed locally. The human features that result from vision processing can be easily compressed for further processing, either locally, or on a server depending on the remaining local resources. The avoidance of video transmission across the network is also required to protect visual privacy.

2.1 The Mobile Energy Crisis

The consumer electronics industry has succeeded in packing large amounts of processing power into small devices. If necessary, custom VLSI can be used to realize virtually any computer vision system on a single chip. Unfortunately, power requirements limit the clock speed and performance of VLSI in mobile applications. Fuel cells may eventually become practical but for now small, wireless devices are fundamentally limited in processing power by battery life and size, and storage capacity and retrieval speed are also limited by energy supply. The impact of energy storage limitations on mobile device performance is reduced by the availability of wireless communication networks that can provide distributed processing and data storage. However, wireless communication (and especially transmission from the wireless device) also consumes power in proportion to bandwidth. Fortunately, a variety of audio, video, graphics, and data compression standards are available to optimize the tradeoff between bandwidth and codec processing requirements.

Wireless vision-based HCI devices must process video from one or more cameras in real-time and deliver the resulting human behavior data stream to an application which performs a desired function. While video compression algorithms have progressed steadily over the years, the delivery of high quality video over wireless networks requires either too much power or too much bandwidth to be practical today. These limitations and the negative effects on vision algorithm performance from video coding artifacts will force the placement of vision processing onto the wireless device. Fortunately, the human behavior data stream is highly compressible and can be transmitted over any wireless network using standards such as MPEG-4 [5, 6].

2.2 User Imaging and Cooperation

An inherent challenge facing vision-based HCI is imaging of the user. The video camera solutions for a particular application environment must address both resolution requirements and camera position and orientation. The continued improvements and reduced cost of CCD and CMOS image sensors has recently made HD video capture available at SD video prices. In particular, the availability of 60 frame per second, HD progressive scan, color video cameras provides new levels of detail and increased field of view. The use of camera pan/tilt controllers further expands the user's freedom of movement.

If performance is more important than cost, size or power consumption, then multiple cameras should be used to provide depth from stereo, reduce occlusions, and reduce feature extraction errors by averaging or outlier removal. Stereo imaging certainly provides better face detection and tracking performance than single camera imaging and face tracking must be reliable enough to perform subsequent individual feature tracking (e.g., eyes and mouth). However, detailed stereo imaging of the face is fundamentally difficult due to a combination of smooth patches (cheeks), holes (nostrils, mouth), and hair; all of which can cause these algorithms to fail. Alternatively, features extracted from each camera can be combined by either averaging or removing outliers. Also note that stereo imaging performance is strongly affected by camera separation which should be optimized for the expected range of the subject. If stereo correspondence is not used, cameras can be placed to accommodate any range of subject motion, or placed in a cluster to increase resolution and/or field of view. Ultimately, a combination of stereo and feature integration can be deployed subject to camera placement constraints.

When automatic user imaging fails, the user can be engaged to either move the HCI device or move her self into view. At this point in the HCI session, the system must present an audio and/or visual display to the user assuming that she intends to interact with the system. In many application scenarios, the content of this display must be understandable to new users with neither experience nor prior intention to use the system. Since people are especially attentive to the human face and voice, the display of a talking human, humanoid, or character is the best way to engage the user in a dialog and optimize her position relative to the camera(s) and microphone(s). The implementation of a talking virtual agent is partitioned into the animation system and the language system. The language system sends animation instructions to the animation system with associated synchronized voice. The stream of animation instructions is inherently low in bit-rate after compression and the animation system is only moderately complex. However, the language system can be very complex and require access to large speech databases and sufficient processing power for real-time response. The best solution for this combination of conditions is, again, a client-server architecture.

2.3 Lighting

The type and position of light sources in the environment obviously directly determine the image signal to noise ratio and variations in appearance of objects in the scene. The degree to which lighting can be controlled or predicted depends on a variety of conditions including user comfort, level of mobility of the HCI device, exposure to sunlight, and the physical/economic practicality of light fixture placement. Non-visible light sources (near infrared) can provide some relief from user comfort issues but must be used cautiously to avoid injury to the retina. Infrared imaging can also be used for some applications.

One's face, body, clothing and accessories are rich with stable color information; at least for some period of time in the case of skin. The ideal camera would be sensitive from infrared to ultraviolet with each pixel expressed as a spectral array of intensities. Image sensors today have non-uniform sensitivity and rely on optical filters to quantize the spectrum. Therefore additional technology with significant additional cost is needed to produce broad-spectrum cameras. The camera is still the cost driver in many applications so the added cost of non-visible imaging may be difficult to justify in consumer applications.

People are quite sensitive to lighting; especially if they are trying to read a screen in a well-lit environment. Diffuse lighting is more comfortable than point sources, and minimizing lighting contrast is also important. Another advantage of diffuse lighting is that shadows are minimized and surface appearance is more stable. A disadvantage of diffuse lighting is that "shape from shading" algorithms are less useful.

2.4 Dialog Systems

The need for a dialog system [7] depends on the predictability of the user's behavior and objectives and also on the degree of user cooperation. For example, at one extreme, no dialog system is needed for vision-based surveillance because the user is not cooperative at all. An example at the other extreme would be an immersive virtual environment with interactive virtual humans or characters. The modes of a given dialog system are chosen based on each modes attributes and weaknesses. The dispersion of the human voice is both useful for broadcast communication and problematic when privacy is desired or noise pollution is a concern. The tactile input mode (keyboard, mouse, touch screen) is tedious for most people but privacy is easier to maintain. The visual capture of human behavior (emotional state, speech, body position) can be accomplished without disturbing others, at a distance, or covertly. However, the reliable capture of arbitrary human behavior in a surveillance environment is still a research frontier. In general, the presentation of audio and visual information to users can be accomplished with the desired degree of privacy; while the user's voice is the most difficult machine input mode to keep private. In addition, the acquisition and recognition of the user's voice is strongly affected by both voice volume and distance to the microphone, compelling users of current systems to speak loudly even if the microphone is close to the mouth. The acoustic input mode suffers from pollution, reliability, and privacy issues. The visual input mode also suffers from reliability issues but should enhance the performance of the acoustic input mode and reduce and possibly eliminate the need for higher voice volume.

The reliable capture of audio/visual user behavior is more easily accomplished when the user is guided and trained and the system can predict when additional guidance and training dialog are needed. While machine understanding of free speech has yet to be fully realized, user speech and emotional state recognition can be used to improve machine understanding of user intent, especially when the user is trained to limit the dialog domain. The achievement of unrestricted dialog between human and machine would be the most convincing demonstration of artificial intelligence. Only a client-server architecture can provide the heavy resources needed to achieve the most advanced dialog systems.

2.5 Privacy and Security

The need for security and user identity verification in all computing and network systems could be satisfied using audio/visual HCI. The rigorous engineering and careful deployment required for any secure system is especially needed with a vision-based system because security is needed for both the visual privacy control and access control subsystems. Fortunately, the real-time acquisition of human behavior data on a local device could provide protection of visual privacy while allowing accurate identity verification over a wireless network by avoiding the transmission of video over the network. The user must be able to reliably control the flow of camera-generated video that is output from the local device. Automatic camera control and video communication systems must be carefully designed to ensure that user cooperation and understanding are maintained. While consumer software companies are not accustomed to lawsuits for malfunction and the typical End User License Agreement (EULA) is notoriously one sided, violations of privacy and security that result from poor design could cause consumer revolt or impede adoption.

2.6 Multi-Model Biometrics

The post 9/11 focus on biometrics-based security has resulted in accelerated deployment of available systems and a government drive to collect biometrics information from as many citizens as possible. The need for identity verification is clear but commercial systems available today suffer from low accuracy, vulnerability to spoofing, or civil rights and privacy issues. For example, a static biometric such as fingerprints can be copied in order to spoof the system. Fingerprints can also be left behind and used to track the past location of people enrolled in the system without their knowledge. Face recognition is not very reliable and is also spoofable. It has the advantage of not requiring physical contact with the user and being socially acceptable. Voice recognition accuracy degrades badly in noisy environments but is difficult to spoof (in quiet conditions) if a challenge response protocol is used (prompting the user for particular utterances). The combination of face and voice recognition and visual speech recognition promises to provide identity verification with much greater accuracy than either mode alone without vulnerability to spoofing. Iris scan is also an option that can be incorporated into access control system applications where lighting and close-view cameras can be used. When the motivation to spoof the system is high and only static biometrics (hand, finger, face, and iris) are collected for unattended access there is a danger of dismemberment by violent criminals. The use of audio/visual biometrics promises to provide accurate identity verification at a distance without endangering the user or violating his privacy. A client-server architecture provides the best protection of user images and voice by secure containment in the HCI device while enabling access control over low bit-rate networks by transmission of compressed audio/visual biometric features (e.g., MPEG-4).

3 Common Application Environments

People need access to information and other people on a moment by moment basis using constantly varying modes that are optimized dynamically. The mobile phone/PDA is currently a handheld voice (and limited video) communicator with less than 50 kilobits per second of reliable bandwidth and adequate audio/visual display. Vision-based user input to mobile phones is currently processor limited but could be implemented in VLSI in the relative near term. Vision-based HCI in vehicles, home, office and public terminals is not constrained by stringent power and size requirements and will be deployed much sooner using commodity components. This section examines how each major application environment presents challenges and opportunities to developers of HCI.

3.1 Mobile

While mobile HCI devices are necessarily handled by the user, fixed HCI devices should interact with the user without requiring physical contact. Busy multitasking people need information and communication systems that work with whatever input modes are practical at the moment. We would all benefit from the option to interact with machines using human-to-human interaction modes (vision and voice) in addition to the traditional modes (tactile). All environments suffer from acoustic noise. This has required the use of close-talking microphones for reliable communication and machine recognition. The integration of visual speech processing into the HCI will bring speech recognition performance up to practical levels for a much greater number of applications without requiring close-talking microphones or elevated voice level. Visual communication with alternate appearance and face/voice recognition for identity verification can also be added as server applications.

3.2 Vehicles

The need for vision-based HCI is greatest for drivers of vehicles given that they are visually occupied while struggling to use tactile interfaces for phone, navigation, and entertainment control. This situation is hazardous enough to compel state lawmakers in a growing number of states to outlaw holding and talking on a cellphone while driving. While voice recognition in vehicles performs poorly due to acoustic noise, audio/visual speech recognition promises to perform reliably enough to be practical. In addition, the recognition of the user's mental state, e.g., fatigue level, using machine vision of head pose and eyelid opening will save lives. Multimodal biometrics applications could also be deployed using face and voice recognition to verify the identity of the driver. Trucks and other large vehicles should be equipped with reliable and convenient driver identification systems.

3.3 Public Terminals

Automatic teller machines (ATMs), vending machines, and grocery store checkout machines are located in public places and currently use simple tactile HCI and a magnetic strip. Public terminals must have robust and minimal tactile interfaces in order to survive dirt, weather, and hostile users. As the use of cash declines and is replaced with electronic payment systems that verify identity the incidence of theft and fraud has increased dramatically. Current credit card security measures do little to foil the determined criminal and electronic identity theft is increasing from already significant levels. Fingerprint readers are highly accurate but could endanger the user or violate his privacy. A major advantage of vision-based HCI for public terminal applications is the ability to complete transactions and verify identity while the user's hands are busy or gloved (no contact required). An advantage of public terminals for vision-based HCI (as opposed to mobile or desk-based locations) is the ability to control the camera placement and possibly the lighting, and model the variations in lighting and view of the users. The use of an animated talking face to engage the user in a dialog should help to reduce the variation in possible user responses. The user can be quickly trained to position herself within view of the camera(s) even if the user was not originally intending to interact with the system. For better or worse, vending machines with talking face dialog systems that beckon to passersby will eventually be deployed.

3.4 Vision-Based HCI for PCs and Game Consoles

The keyboard and mouse continue as the HCI of choice for personal computers in spite of the availability of speech recognition systems that require close-talking microphones for sufficient accuracy. Low typing speed and repetitive strain injuries are still preferred over state-of-the-art speech recognition systems. While CPU speeds have increased on schedule, the processing needs of vision algorithms still consume most or all of the latest PCs power. Just as graphics acceleration hardware became standard equipment on PCs to free the CPU for other applications, vision acceleration hardware will eventually become a standard for user identity verification, visual speech recognition, user state and gesture recognition. No head-mounted microphone will be required to interact with the PC using speech recognition. Gaze tracking will be used for spatial selection and mental state recognition, and gesture recognition will eventually become practical.

The deployment of vision-based HCI in the home requires that visual privacy be controlled in close cooperation with the user. Once the images from a camera are stored in a computer memory or disk, they are vulnerable to malicious or inadvertent transmission over the Internet by viruses or novice users. The vision-based HCI peripheral should be able to extract human behavior data from the video and transmit it to the PC without transmitting the video itself. This visually private operating state should be clear to the user and not changeable remotely. Consumers will need to learn to trust such systems before they are widely adopted.

4 Current and Emerging System Examples

So far, this chapter has analyzed the architectural requirements and environmental constraints that should inform the design and deployment of practical vision-based HCI systems. Given the small number of these systems in the field today, this analysis has been largely theoretical and somewhat speculative. This section describes commercial systems that are either available to consumers or employees now, or could be available now if the market were ready. Systems that involve direct contact with a sensor (e.g., fingerprint readers) or very close viewing and restricted user movement (e.g., iris scan) are not covered here.

4.1 EyeToy

Recently, a vision-based game controller called EyeToy [8] was successfully introduced to the consumer market by Sony for the PlayStation2 with games specifically designed to incorporate real-time imaging of the user. PS2 inputs video from the EyeToy camera via USB and performs all vision functions using the standard PS2 computing resources. A typical EyeToy game tracks gross body and arm movements in real-time and provides the user with visual feedback using overlay graphics on video of the user. As the first mass deployment of vision-based HCI to consumers, the evolution of EyeToy will be interesting to watch. Figure 1 shows the EyeToy in action from the gamers point of view as he attempts to bounce the virtual soccer ball off of his head. Special colored props can also be tracked by the system.

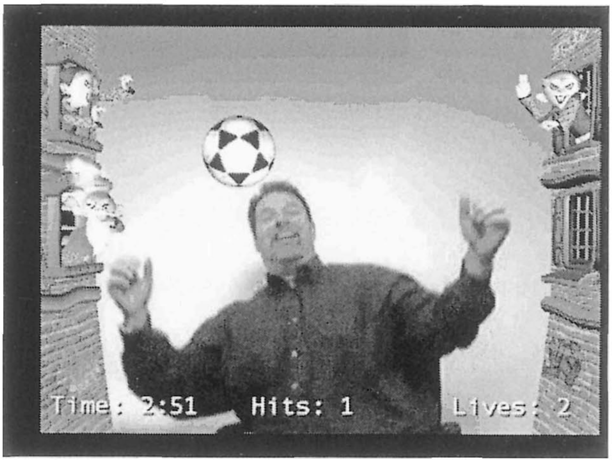

Fig. 1. Sony PlayStation EyeToy screen shot (courtesy R. Marks, Sony Computer Entertainment US)

4.2 Driver Eye Tracking

Vehicle driver face and eye tracking has not yet been commercially deployed but the technology has reached a level of maturity that makes it practical for prevention of falling asleep while driving. Cost and user resistance are the main barriers to deployment in the future. An example of a single camera system has been developed by the Delphi Corporation [9], and a stereo vision system has been developed by Seeing Machines [10].

4.3 Access Control Systems

The average consumer has to deal with several key or code based access control systems for use of vehicles, ATMs, credit cards, cellphones, computers, web sites, and buildings. These systems are not very secure as evidenced by the high rates of auto theft, credit card fraud, cellphone fraud, computer viruses, lock picking, key theft and duplication, etc. Biometrics access control technology [11] is available for much better security but cost and social issues are still holding back widespread deployment except for their mandated use at international border crossings since 9/11. Cost will continue to decline but the social and political issues could intensify as the need for security increases and personal privacy is challenged. Consumer adoption of biometric security for public terminals (ATMs, gambling and vending machines) will be limited by resistance to the enrollment process where the user provides proof of claimed identity and submits to the collection of biometrics. While everyone is affected by the cost of theft and fraud, the financial institutions, casinos, and vending machine companies have the greatest incentive to improve security using biometrics. Consumers will probably need additional incentives to cooperate; especially if fingerprints or other problematic biometrics are collected.

Biometric access control systems are currently being deployed in the workplace and in airports for both international passengers and workers. The US Visit program requires face and fingerprint biometrics to be used to verify the identity of Visa holders wishing to enter the US with the program expanding to all passport holders but the system is attended by customs and immigration agents. Schiphol Airport in Amsterdam has deployed iris recognition systems for automatic access control for volunteer passengers and airport employees [12]. Face recognition by itself has not been significantly deployed for access control [13] in the workplace or transportation systems.

The use of face recognition to access PCs, game consoles, and secure Internet locations can be deployed as a local application by the user as a replacement for passwords. A variety of systems are commercially available [14] but not widely used. Perhaps vision-based access control would be adopted more widely on PCs if other vision-based HCI applications were also deployed.

4.4 Immersive Simulation

The military is the leading developer of reality simulation systems with the "human in the loop." Head and eye tracking systems are currently deployed in many of these systems [15] as part of the HCI and for measurement of human performance. As real-time facial capture from video and audio/visual speech recognition systems mature, the emotional state, speech, and gestures of the user will also be available for simulation applications. The successful use of a complete vision-based HCI system in simulation should be rapidly followed by cost reduction, miniaturization, and ruggedization for deployment in vehicles, command and control centers, and finally mobile devices.

5 Conclusions

The adoption of technology by consumers is the ultimate validation of its maturity and utility. Vision-based HCI related applications have barely started to penetrate the consumer market and industrial deployment is mostly limited to access control systems where the period of use is inherently very brief. The potential for wide deployment of vision-based HCI is great; especially in applications where speech recognition is also needed. In particular, the rapid consumer adoption of advanced mobile phones [16] with video cameras promises to provide a platform for vision-based HCI using a client-server architecture and standards like MPEG-4 FBA. Rapid adoption is also possible on PCs, game consoles, and vehicles when enough vision-based HCI applications are available to justify the cost.

Acknowledgments

Thanks to the editors of this book for giving me the opportunity to express my views on the present state of vision-based HCI applications and the future paths to widespread adoption.

References

1. J Segen and S Kumar. Gesture VR: Vision-based 3D hand interface for spatial interaction. *Proc ACM Int Conf on Multimedia,* 1998.
2. E D Petajan. Automatic lipreading to enhance speech recognition. *Proc CVPR*, 1985.
3. A Goldschen et al. Continuous optical automatic speech recognition. *Proc Asilomar Conf on Signals, Systems, and Computers,* 1994.
4. G Potamianos et al. Large-vocabulary audio-visual speech recognition by machines and humans. *Proc Eurospeech*, 2001.
5. ISO/IEC 14496-1 IS (MPEG-4). Information Technology – Coding of audio-visual objects, Part 1: Systems. `www.iso.org`
6. ISO/IEC 14496-2 IS (MPEG-4). Information Technology – Coding of audio-visual objects, Part 2: Visual. `www.iso.org`
7. R Cole et al. Perceptive animated interfaces: First steps toward a new paradigm for human-computer interaction. *Proceedings of the IEEE.* pp 1391–1405, 2003.
8. R Marks. Natural interfaces via real-time video. *SIGGRAPH 2000 Sketch.* `research.scea.com/research/pdfs/siggraph2000RICKnat_interfaces.pdf`
9. B Kisačanin et al. Driver Drowsiness Monitor from DELPHI. *Proc CVPR Demonstrations (CD-ROM)*, 2004.
10. `www.seeingmachines.com`
11. `www.biometrics.org/html/examples/examples.html`
12. `www.biometritech.com/features/deploywp1.htm`
13. `www.cisco.com/en/US/about/ac123/ac147/archived_issues/ipj_7-1/lures_of_biometrics.html`
14. `www.biomet.org/faceproducts.html`
15. `www.hf.faa.gov/docs/508/docs/VF-SNIPVFRDarken.pdf`
16. `www.mobilepipeline.com/59200081`