15

User Interaction for Mobile Devices

Sanni Siltanen[1], Charles Woodward[1], Seppo Valli[1], Petri Honkamaa[1], and Andreas Rauber[2]

[1] VTT Technical Research Centre of Finland
[2] Institute of Software Technology and Interactive Systems

Smart phones have the potential to become the default physical user interface for ubiquitous mobile multimedia computing applications [40]. However, the conventional tiny keypad of a mobile phone is unsuitable for many situations; for example, typing a simple URL like http://www.google.com/ might require over 70 key presses. The fast development of mobile devices provides new possibilities for user interaction methods. Accessing large amounts of multimedia information requires specialized methods for searching and data mining. Besides direct interaction with keyboard, keypad, joystick, touchscreen or even speech recognition, interaction with mobile devices can be enhanced by analyzing the location or motion of the device. Furthermore, it is possible to fine tune the location information with the device's direction and orientation information using visual markers or various sensors. In addition, the movement of the device can be analyzed using accelerometers or analyzing optical flow of the attached camera. Augmented reality is a related technology, providing new interaction possibilities with a visual link between physical and virtual worlds.

15.1 Evolution of Mobile Multimedia

Multimedia applications have existed as long as there have been personal computers supporting the playback of audiovisual (AV) content. Especially after PCs were equipped with a network access for downloading content, the development of multimedia applications has been very rapid. After the emergence of the Internet and the world-wide-web, multimedia content has evolved towards more and more rich content including various kinds of textual, aural, and visual content.

Multimedia content includes, e.g., images, audio, music, video, films, news, documentaries, advertisements. The application types can be classified in two main categories, namely local and networked/streamed applications. PC games are a typical example of local applications, whereas video conferencing,

P. Maragos et al. (eds.), *Multimodal Processing and Interaction*,
DOI: 10.1007/978-0-387-76316-3_15, © Springer Science+Business Media, LLC 2008

e-commerce, e-learning, IP-telephony and IP-TV are examples of streaming applications. Various data formats and standards are being used for multimedia content, e.g., MP3 for music, MPEG4 for video, and various W3C specifications for web content.

During the time mobile phones have been on the market, various PC based applications have been ported into mobile platform, as far as it has been technically feasible. In recent years, the processing power, amount of memory, multimedia capabilities, and access speeds of mobile phones have approached that of PCs, which has made this transition easier and faster. Thus, the typical office applications, i.e., email, calendar, document viewers, editors, etc. have successfully been ported to mobile environment. Another important category are web applications, based on new mobile browsers and the increased bandwidth available for wireless data transmission.

In addition, there are several mobile applications which take the characteristics of a mobile terminal and user into account. A good example of this is SMS, which has not a direct predecessor in the PC environment. Following SMS, ring-tone downloads and image screen-savers have been surprise successes for mobile phones. As in the PC environment, local applications have been easier to implement on mobile devices, although the development goes towards real-time streaming applications.

As the display and keyboard of a mobile phone are small, mobile multimedia applications meet particular challenges for the supported content types, usability and interaction. The interaction in mobile applications has long been based on miniature versions of a mouse, joystick, and keyboard. Today, the mobile phone's camera is also increasingly used for interaction, e.g., for pointing and tracking (some mobile phones even provide an integrated accelerometer that can be used for the same purpose).

A specific opportunity in mobile applications, particularly challenged by usability issues, is providing authoring and editing functionalities for multimedia content, e.g., for either downloaded or user generated video clips. Recently, the number of location and community based mobile applications has been rapidly rising, supported for example by the integration of mobile phones with a camera and a locationing device (GPS).

A comprehensive collection of various mobile multimedia related research activities is given in [221].

15.2 Mobile Multimedia Terminals

The prerequisites for mobile multimedia solutions are the availability of: (i) suitable, e.g., powerful enough mobile terminals, (ii) intelligent, fast and reliable software, and (iii) applicable networks, i.e., modes for communication. Considerable amount of memory and computational power is required especially for capturing and processing (e.g., compressing) of images or video,

displaying of 3D graphics, as well as for managing the transmission and connectivity, e.g., protocols and algorithms to cope with roaming.

Today, mobile phones, PDAs (Personal Digital Assistant) and other mobile devices are already quite capable to see, hear and sense the surrounding environment by various means. Mobile phones and other hand-held devices are advancing towards powerful communicators with multiple network access (GSM/GPRS, 3G, WLAN, Bluetooth), multimedia capabilities (digital cameras, high-resolution color displays, MMS), open platforms for applications (Java, Symbian, etc.) as well as for Internet connectivity.

The classification between the devices is becoming vague as they are providing more and more common features; for example PDAs providing in-built mobile network access and camera, and mobile phones providing larger displays and more advanced user interaction mechanisms, e.g., keyboards and input pens. Therefore we can think that in the future the different mobile device categories will merge into a *smart mobile terminal* containing all the needed functionality.

Miniaturization of tablet PCs has led to a new class of devices, nowadays called Ultra Mobile PCs (UMPC). As example of UMPCs, the recent Sony Vaio UX models is equipped with a slide-in keyboard, two built-in cameras, wireless Bluetooth, WLAN and 3G connectivity. Also, PDAs are available in a variety of models, with various operating systems, with adequate processing power, battery life, and display capabilities for mobile multimedia.

The latest smart phones, e.g., Nokia N95 and E90, have more processing power and even better 3D graphics support than super computers used to have ten or fifteen years ago. Among the different manufacturers, Samsung has been particularly active to provide the phones with new integrated functionality, such as high accuracy cameras, accelerometers and other sensors. Apple iPhone is leading the way for new interaction methods on mobile phones, with multi-touch screens and 3D visual browsing techniques.

15.3 Mobile Displays

The screens of hand held devices are obviously the most common displays for mobile multimedia applications; however there are also other options available. Besides wearable devices such as video glasses, also projector displays are developed as display extensions for mobile devices [338].

The application specifies which features are critical and the application type defines the best display type for it. On mobile phones factors, such as the small display size and resolution have to be accounted for, while on PDA and UMPC devices the screen brightness and power consumption can be more critical factors. Handheld mobile devices are generally preferred if the task does not require hands-free operation or immersive display; otherwise, head mounted or projected displays can be a better choice [64].

15.3.1 Stereo Displays

Conventional displays are monographic, but various methods are available to upgrade even standard displays to reproduce stereographic view. The standard methods can be divided into two groups named active and passive stereo. Both active and passive stereo require that the user wears special glasses meant for this purpose. With active stereo the display alternates the image meant for left and right eyes rapidly so the shutter glasses show the right image for each eye. With passive stereo, both images are shown at the same time and the glasses filter the correct image for each eye, e.g., red-green-glasses.

The third group of stereographic displays are called autostereoscopic displays that do not require any separate glasses to be worn. They usually show the left and right images for separate viewing segments, either by tracking where the viewer is located or just by offering several discrete viewing segments. Samsung has already demonstrated such stereo displays on smart phones. The primary application for stereo displays on camera phones would be mobile 3D games.

At the software level, StereoGames [576] is a solution based on anaglyph technology to enable changing originally monographic 3D applications to stereo. Besides PCs and game consoles, StereoGames works also for mobile devices like mobile phones, and it can be applied to create stereo effect for both passive and active as well as autostereoscopic displays.

15.3.2 Head Mounted Displays

Head mounted displays (HMDs) are wearable display devices ranging from helmets providing deep immersion and full field-of-view, down to miniaturized data glasses or "goggles". There are two types of HMDs, based either on video or optical see-through approach [64].

The most popular multimedia application for HMDs is watching movies on portable MP3 and DVD players. Other uses for HMDs are found in mobile games (using video glasses), and in augmented reality (using either video or optical see-through approach). Recently one of the leading HMD display companies MicroOptical changed name to MyVu and is now providing video glasses for the iPod [352].

Some visionaries and researchers are developing concepts and prototypes where the reproduced image is not shown on any display; instead the image is drawn directly to the user's retina. Thus, the technology is called virtual retinal display (VRD). However, commercial VRD products are not yet available.

15.4 Interaction Modalities

According to Foley et al. [167], input devices can be categorized by the graphics subtasks they can perform. Those tasks are: position, orient, select, path,

quantify and text entry. The ability to perform these tasks and the ability to interact with the environment are essential for future ubiquitous multimedia devices. In the following, we explore various interaction modalities, starting from various methods available for text entry, up to a discussion on motion estimation for gestural interaction on camera phones.

15.4.1 Text Entry: Keyboards, Strokes and Dynamic Selection

Text entry methods with mobile devices can be divided broadly into following categories: keyboards, gestural alphabets and dynamic selection techniques. Next, we describe briefly some popular text entry methods. More detailed summaries can be found in [561, 315].

Currently the most common input method with a 12-key telephone keypad is multitap, where the user presses each key one or more times to write each letter. The 12-key input can be optimized by adding language knowledge to the system. One example is the T9 solution licensed by many phone manufacturers [109]. In T9, the keys are pressed only once for each letter and the linguistic model predicts the most probable word for the key sequence.

Touchscreen displays, e.g., on PDAs and some multimedia phones, provide the option of a screen keyboard. Some other devices, e.g. ultra mobile PCs, provide a larger keyboard that is normally hidden but can be slid or flipped out when needed. Alternatively, a full size keyboard can be a separate accessory connected using, e.g., Bluetooth. A novel approach for fitting a full QWERTY keyboard on a small device is to project the keyboard on a table with a laser beam and recognize the keyboard taps optically [309].

Keyboards and screen keyboards may also be rearranged, either to fit into smaller space or to provide faster interaction. The number of keys can be reduced either by allowing toggle operations between different letter sets, or by using key-combinations, a.k.a. chords. A good example of reduced key set is half-QWERTY keyboard which allows writing speeds up to 73% compared to writing with a full keyboard [315]. Another example is the Twiddler chord keyboard [110] which is particularly popular with wearable computing researchers.

Free-form handwriting recognition is a widely available, though not very reliable text entry method on PDA devices. An alternative to normal handwriting is the Graffiti system on Palm devices, in which each letter is written using a single stroke that is similar enough with normal hand-written characters. Another approach for gesture based writing is Quikwriting introduced in 1998. It is based on 3x3 grid where the strokes are started and ended in the center; shortest strokes need to visit only one grid position while more infrequent letters need to make a curve through several grid positions [397].

In dynamic selection techniques, the shown alphabet is changed dynamically. For example, in FOCL (Fluctuating Optimal Character Layout), the choices for the most probable next letters are always shown near the center [51]. In Dasher [561], the possible first letters are shown in a column of boxes

where the box sizes reflect the letter frequency; within each of these boxes the possible next letters are presented in a similar box column, and typing is performed by moving a pointer to continuously zoom deeper and deeper in the nested boxes.

15.4.2 Joysticks

Besides the traditional keyboard or keypad, current mobile devices provide various other ways for direct interaction including joystick, trackball, touchpad, and camera. Practically all mobile phones today include four- or eight-direction navigation joysticks. Usually these navigation buttons have only binary resolution but there has also been some research using isometric joysticks [489].

15.4.3 Multi-Touch Screens

Many recent mobile devices provide a touchscreen with a natural pointing and drawing user interface. Apple has developed this idea further with the iPhone by providing a multi-touch interface. The iPhone interface is based on a capacitive touch panel and it can recognize several gestures with one or multiple touch points. The gestures are used creatively in applications for example for scaling photos, zooming in maps or for flipping through music albums or photos. The iPhone incorporates also an intelligent touchscreen keyboard that compensates the small keyboard key-size by using dictionary for favoring the more probable keys and correcting spelling mistakes on the run.

Multi-touch interfaces are currently a hot research topic as they are believed to provide more natural and more versatile interaction possibilities compared to single-touch displays. In the "non-mobile" world, interesting interaction possibilities have been presented for example by Jeff Han [195]. Also Microsoft is releasing a commercial Microsoft Surface product based on multi-touch [337]. Besides multi-touch, the upcoming Microsoft surface promises also a natural interaction with mobile devices. One presented example application includes moving photographs from your digital camera into your cell phone. In the example, this is done just by placing both devices on the multi-touch surface and dragging the photos between them.

15.4.4 Haptic Systems and Tactile Feedback

Haptic systems often refer to virtual reality input/output devices where the user can not only manipulate virtual objects directly, but also gain tactile feedback of how the virtual object feels like. With mobile devices, the most common haptic feedback is given using vibration actuator common in many phone models. This can be used for example to produce a "poor man's force feedback" in mobile games.

Several more advanced haptic systems have been presented, though less suitable for mobile use. Most typically the sense of touch is created by using mechanical or pneumatic construction, e.g., [475]. Some other methods do not restrict the motion of the user's fingers and hands, but instead they generate vibrations, e.g., using array of pins against fingertip.

15.4.5 Gesture Based Interaction

Physical manipulation of the handheld device is an integral part of an embodied user interface [162]. This may involve, e.g., moving or shaking the device, leading to interaction paradigms such as "squeeze me, hold me, tilt me" [198]. Gestures by tilting or moving the device may be used in positioning and pointing tasks or to indicate commands. For example, the user could scroll a menu on the phone's display by tilting the phone up- and downwards, select the command by shaking the phone to the left, and get to previous menu by shaking to the right.

Various sensing techniques already exist for mobile interaction, including accelerometers, touch sensors, proximity sensors, and pressure or squeezing sensors [211, 162]. In addition, the orientation of the mobile device can be analyzed in relation to earth gravity using tilt sensors, or in relation to earth magnetic field using compass. However, such sensors are seldom available as standard components on phones or other mobile devices; instead they have to be attached to the device as bulky accessories, which ordinary users have seldom available. Although this situation may change in the future (accelerometers are already integrated in some mobile phones), the most common "sensing" accessory on smart phones for the next few years will still be the camera.

15.4.6 Methods for Camera Motion Estimation

Solutions using camera image based tracking span from simple motion tracking implementations up to 3D feature tracking. The most sophisticated feature detection and tracking algorithms are able to derive 3D coordinates of the physical world from the camera view, e.g., SLAM (Simultaneous Localization and Mapping) [124]. In theory, the camera's optical flow gives enough information for 3D reconstruction of the scene and the 6 degrees-of-freedom motion path within that scene [199].

For interaction purposes, the global optical flow motion in the mobile device's camera view can be used to approximate the device's movement. Motion estimation is typically implemented by searching image blocks in various displacements within the previous image frame. As an exhaustive search is usually too tedious, the motion search process typically has to be simplified on mobile devices.

For example, the TinyMotion algorithm [558] reduces the computation complexity by using grid sampling (a multi-resolution sampling technique) for the input image before making full-search block matching algorithm. The

Projection Shift Analysis (PSA) [142] approximates 2D motion by using horizontal and vertical projection buffers of the camera image instead of image blocks.

The complexity can be reduced by tracking only "easy-to-track" features of the image. For example, Hannuksela et al. [196] propose using pixels having maximum squared differences compared to adjacent pixels. Another common feature selection method is to track corner features in the image, which in turn are found by Harris or SUSAN corner detection algorithms, or by eigenvalue methods [482].

Usually for all vision based motion or feature detection methods, horizontal and vertical camera movements are much easier to analyze than motion in the camera's depth direction; the reason being that there is not much change in the image when the camera is moved back or forth. Modern twin camera phones offer a convenient solution to partly overcome this problem: depth motion can be analyzed from the camera facing the user, having the user's face as a close target for depth comparisons.

In order to solve the motion detection problem accurately, flexibly and fast enough, a hybrid method, i.e., combination of different algorithms and sensors, is often required. A typical hybrid solution is to apply vision based camera tracking while the camera is relatively stationary, and accelerometers during fast camera movement.

15.4.7 Further Interaction Modalities

Instead of keyboard, touchpad, joystick or gestures, mobile interaction may be based also on gaze tracking, or even on user's breath [223]. These more exotic ways of interacting are suitable especially for disabled users. Spoken dialogue and multimodal interfaces are also important modes of text entry for mobile phones.

15.5 Context Aware Applications

15.5.1 Mobile Context Categories

Research for context-aware mobile applications has been reported since the early '90s. Context can be divided into four categories: computing context, user context, physical context and time context [466]. Computing context contains, e.g., network connectivity, communication bandwidth and nearby computing resources. User context includes information like user's preferences, current location and nearby people. Physical context describes physical conditions like lighting, noise levels, and temperature. Time context defines the date and time based information. Besides using the current context information, the context history is also useful for some applications. Further, context-aware applications may actively adapt their behavior based on the discovered

context (active context), or leave the decision on its use to the user (passive context) [92]. Context information may be used in various ways. For example, the application font size can be made larger when the user is walking, and the mobile phone ring volume and vibrate option may be adjusted depending on the situation. One important application area for context aware services is giving guidance for the user. The location information may be used, e.g., to inform the user about nearby services, suggest routes or give more information about observed attractions. The museum visitors may be given personalized TV-like presentations depending on their location, their facing direction, device orientation and their interests during the visit [439].

Various other concepts, definitions and classifications for context exist. Next, in Section 15.5.2 we consider context aware interaction with spatially bound information. Especially, we discuss physical browsing, where the context is determined using nearby electronic or visual tags. For more thorough surveys on context-aware services see [92].

15.5.2 Spatial Information and Interaction: Physical Browsing

Spatial information consists of the physical location, orientation, and the information associated (or bound) to the respective visual view. The location can be derived either manually, by using satellites (GPS, Galileo), mobile network positioning, local network (WLAN) based methods, or short-range methods (Bluetooth, RFID , NFC , etc.). More advanced functionality can be built when taking also orientation (viewing direction) into consideration; the orientation may be derived, e.g., by electronic compasses or cameras.

In applications using spatially bound information, means are needed to search for, discover, and browse information in the environment. This is generally referred to as physical browsing, which also includes optical/camera based methods. In physical browsing, tags are typically used to provide the user with access points for available information or services. A variety of tag types are used: visual tags (e.g., barcodes and matrix codes), RFID tags (Radio Frequency Identifier), NFC (Near Field Communication) tags, Bluetooth tags, Infrared tags, etc. [18]. RFID/NFC tags typically communicate over relatively short distances.

Usually electronic tags are indicated by some visual sign as well. Pointing is perhaps the most natural user interface, and therefore the basic user paradigm in physical browsing. Children use pointing inherently in all cultures even before learning to speak. Remote controllers for home appliances are typical pointing devices at home. In addition to “point me”, other user paradigms introduced in physical browsing are “sweep me”, and “touch me” [406].

When noticing an access point, the user typically points at the tag with his/her mobile device to see the available information and/or make the desired action. A tag itself, either electronic or visual, may contain encoded/stored information and metadata telling what to do with the information, e.g., this is a phone number, make a phone call; this is a web address, start browser.

The user may for example establish a phone call just by pointing at a tag attached to the picture of a person [491]. A tag-based user interface can also be built to support multimodal interactions [252].

A further benefit of visual tags is that they can be used to derive additional information about the movements of the camera, including tilt, rotation, and distance in relation to the tag. Thus using visual tags for context-aware mobile applications can support free augmentation of the physical environment with spatially bound information, including efficient functionality for information detection, interaction, authoring and sharing [190].

Drag-and-drop is a paradigm where the user clicks a virtual object and drags it to different location or to another virtual object. Related to the drag-and-drop concept is the term hyper-dragging, where the user can transfer information from one computer to another (or from a mobile device to a computer), by only knowing the physical relationship between them [435].

15.6 Mobile Augmented Reality

Augmented reality (AR) is a relatively new concept within computer graphics and video processing research, yet with high potential of becoming an integral part of future mobile multimedia interfaces and services. The 2007 MIT Technology Review [241] lists mobile augmented reality (MAR) as one of the ten technologies "most likely to alter industries, fields of research, and even the way we live".

Basically, augmented reality means superimposing digital objects into the user's view of the environment [36]. The real world and a totally virtual representation are the two ends of the Mixed Reality (MR) continuum [339]; augmented reality is situated in the middle of this continuum. Besides 3D presentations, simple graphics elements such as augmented text and symbols can be applied for providing guidelines and additional information to the physical world. Instead of static content on the mobile terminal, in the future, augmented content will be increasingly provided by ubiquitous connections to Internet and local services.

The potential of wearable augmented reality has been investigated at the early stages of AR research, but until recently wearable AR systems have often been too heavy and resource demanding for practical applications. Today, the rapid development of mobile devices has lead to small devices with enough processing capacity, 3D graphics support, high resolution displays, built-in cameras and long lasting batteries to enable light-weight mobile AR systems [194]. Some examples of how tablet PCs, UMPCs, PDAs and camera phones have been successfully used in various mobile AR applications are provided by [390, 214, 207, 441]. Fig. 15.1 shows a lightweight augmented reality system [213] running on camera phone.

The most challenging task for mobile augmented reality is tracking, i.e., accurate and fast mapping of coordinates between physical and 3D virtual

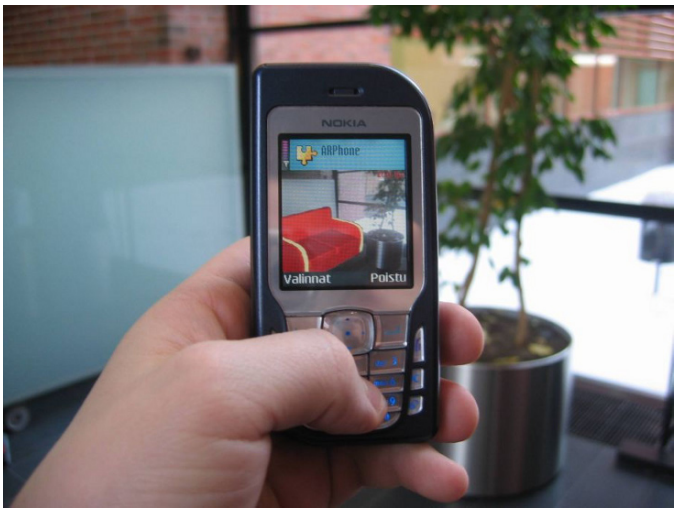

Fig. 15.1. Example of augmented reality on mobile phone: virtual sofa in real environment.

worlds. The mapping is most typically done using information acquired with the mobile device's video camera. Tracking with stationary AR applications is most typically based on using visual markers; this approach can be used also with camera phones as long as the user does not move around too much [207, 441]. However, markers are generally too restrictive for mobile applications and markerless solutions are called for.

Markerless augmented reality is typically based on 3D feature tracking, using methods such as mentioned in Section 15.4.6. Additionally, the augmenting system has to be initialized with some a priori information of the view, for example, in the form of visual markers or beacons, or by having a virtual model of the scene as reference. The augmentation can also be simplified by making the process semi-automatic; using manual interaction in tasks that would be complex to automate but are easy for the user and accurate enough for the application.

Another approach for simplifying the augmenting task is to use just still images for augmenting. This enables for example the use of more sophisticated 3D mapping/tracking algorithms as speed is not a critical factor; also if required some parts of the computation and/or rendering can then be offloaded to a server machine. In many applications, the still image principle works actually more ergonomically than keeping the augmented information in real time video view. In addition, still images provide better image resolution and thus improved accuracy for augmented reality.

A good introduction to augmented reality interaction on camera phones is provided by Henrysson et al. [207]. They describe and compare several marker-based methods for interactive 3D object manipulation, e.g., selection,

translation and rotation. Evaluation of markerless augmented reality interaction on mobile phones is an important topic for future research.

Different users have different preferences regarding how to interact with an augmented reality system. For example, a demo system to test multimodal user interface in AR aided assembly task is presented by [490]. This system has three input modalities (traditional keyboard, speech and gesture control) and visual feedback (output) modality. Test users favored a multimodal input interface and desired more feedback on the output side.

15.7 Example Applications

The following sections give examples of some new generation mobile multimedia applications. Gesture based interaction is discussed in Section 15.7.1, and outdoors augmented reality applications in Section 15.7.2. Section 15.7.3 presents an application of map-based music content interfaces.

15.7.1 Gesture Based Applications

SymBall [193] is a virtual table tennis game on camera phones, using the phone's movement as the sole interaction method to control the game. The movement of the phone is detected using the optical flow of the camera, which is transformed to the movement of the virtual racket in the game. The user can adjust sounds, ball speed, racket shape, etc. for the game. The game can be played in single player mode against "the wall/machine" or against another player over Bluetooth connection. The implementation has also been extended to use GPRS/3G connections. Furthermore, the game also exists as stereo version; see Fig. 15.2.

Gesture based interfaces can also be used to control external devices, for example using Bluetooth connection with mobile phones. The Phonecam-based sweep technique [41] is used to interact with large displays, making the phone act as optical mouse. Similarly, the PhoneMouse software [552] makes the camera phone work as an optical mouse for a PC: moving the phone in the air moves the cursor on PC screen, while the phone keys simulate mouse buttons and launch actions such drawing. See Fig. 15.3.

Viewing of panoramic images is a further example of multimedia applications that can be implemented intuitively based on the camera's motion detection: turning the camera phone around shows views in the panorama image accordingly. Applications for this include presentations of real estate and apartments, display of public places and related information, as well as viewing of 3D virtual architectural models of past and future.

15.7.2 Outdoors Augmented Reality

One of the earliest mobile AR implementations is the Tinmith backpack system, which since its introduction in 1998 has undergone various developments

Fig. 15.2. SymBall virtual table tennis game on camera phones[193], using StereoGames software [576] to create depth illusion.

Fig. 15.3. Using PhoneMouse [552] to annotate a Powerpoint presentation over Bluetooth connection.

in both performance and size [374]. Tinmith includes various hardware devices such as immersive HMD display, GPS locationing, compass and gyrometer. As a special feature, Tinmth provides also the possibility for user interaction with the augmented data using data gloves. A well-known application example with the Tinmith system is mobile real-life implementation of the ARQuake game.

In a more recent architectural application, AROnSite [214], the virtual building's scale and orientation are deduced automatically based on its placement in Google Earth and the user's GPS coordinates. However the actual placement to the scene is done interactively by visually aligning the virtual object to the scenery. After the manual initialization, the camera's optical flow is analyzed to keep the augmented object in place. See Fig. 15.4.

Fig. 15.4. Virtual building added to Google Earth, and visualized on site with a mobile device using AROnsite [214].

.

An impressive hybrid tracking approach for outdoor mobile augmented reality is presented by Reitmayr et al. [434]. Their approach uses an edge-based tracker for matching the scenery with a coarse textured 3D model of the existing environment. Very good accuracy and robustness is obtained by using an additional sensor pack providing gyroscopic measurements and 3D magnetic field vector.

On mobile phones, Nokia has developed the MARA prototype system for sensor based mobile augmented reality [371]. The system consists of Nokia S60 platform phone and attached external sensor box with accelerometer, tilt compensated compass and GPS providing position and orientation information to the phone via Bluetooth connection. The application augments the camera

view with graphics and text in real time, annotating the user's surroundings with location-based Internet content, e.g., tourist information.

15.7.3 Map-Based Audio Retrieval Applications

Since digital music collections are growing constantly, especially on mobile devices, it is getting more and more important to provide easy and intuitive access to these collections. Today it is possible to have thousands of songs on mobile devices but we still miss more adequate ways of accessing music than merely scrolling through directories or hierarchical structures. A possible solution to this is to use a *music map*.

To create *music maps*, first of all a feature extraction algorithm is used, in order to automatically extract semantic descriptors from the audio, which form the basis of determining similarities between pieces of music. Afterward, different clustering algorithms, e.g., a Self-Organizing Map, can be applied to the extracted feature vectors, in order to create a representation of a music collection on a map. The basic idea of clustering is the identification of coherent sub-groups of similar instances, i.e., pieces of audio. As a consequence, pieces of music will be represented on the resulting map and grouped according to acoustic similarity.

User interfaces for music based on the SOM have been researched by several teams, resulting in miscellaneous applications on the desktop [343, 265]. For mobile devices, however, interfaces have to fulfill special requirements. In addition to the size of the display itself, input possibilities for mobile devices are heavily restricted. The most difficult question, however, is how to adequately display and provide access to a large-scale dataset using these minimal interface capabilities.

Fig. 15.5. The PocketSOM application running on a Nokia 7710 emulator.

PocketSOM is a light-weight version of the *PlaySOM* application (an overview of which is given in Section 11.2), which provides a simple but intuitive interface. The interface presents a graphical landscape with the metaphor

of islands in the sea (*"Islands of Music"*). The islands represent the clusters, where music which is acoustically very similar (from the perspective of the features extracted from it) has been aggregated together. In the sea in-between the islands, music that is less strongly represented can be found, with smooth transitions in terms of genre or style from one cluster (island) to another. These music maps are generated by the *PlaySOM* application and then exported to the *PocketSOM* on the mobile device.

PocketSOM's functionality focuses on the interaction of the user with the music map to directly retrieve music and create playlists. Using the touchscreen the user can draw a path on the map, which will result in a playlist. By drawing a path from one island (cluster) to another, smooth transitions from one musical style (or genre) to another can be generated. The resulting playlist is presented to the user who can refine it by re-sorting or deleting titles, and then play it back in different ways: The music can be instantly played on the device with an audio player (if the music is stored on the device) or streamed via an Internet connection from a remote server. Alternatively, the playlist can be also exported, and opened later with a player either locally or on another device. As another possibility, PocketSOM can be used also as a remote control by sending the playlist to a player on another device, e.g., having a PC with the entire music collection playing it back.

As a conclusion, PocketSOM offers an intuitive and convenient alternative to traditional music selection by browsing and constitutes a new model of how to access a music collection on portable audio players.

15.8 Conclusions and Future Directions

Future multimedia applications and systems will evolve towards more ubiquitous, user and situation (i.e., context) aware, adaptive, proactive, and intelligent solutions. The trend is towards embedded wearable mobile computers with new displays for mobile and ubiquitous multimedia. Solutions for the perception, sensing, and modeling of the environment will be increasingly important.

As another main trend, user created content has begun to play an increasingly important role in mobile multimedia. Application platforms such as Microsoft's Photosynth will offer means for massively multi-user content creation, by augmenting of virtual worlds with photos of the real one, and fusing these into new representations such as panorama views and 3D reconstructions. Such alternative representations will in turn promote both new multimedia content and applications for augmented reality and related ubiquitous solutions.

New client-server solutions are needed taking also security and rights management issues into consideration. Reaching "beyond the SMS era" in mobile applications still requires work on new service platforms and terminal applications, especially to demonstrate and evaluate new mobile multimedia com-

munity service concepts. Software components and prototypes for enriching, managing and provisioning of mobile content must be demonstrated and promoted for these communities. User experiences and evolving usage cultures must also be analyzed, as well as suitable business models for community-based services.

Overall usability is a particularly important issue in mobile multimedia applications. This is mostly due to the restricted size of a mobile device, especially the small display and keypad. In addition, natural and intuitive gestures vary in people with different backgrounds. A lot of work is still needed to overcome these difficulties. This work includes research in human-computer interaction (HCI) technologies, user-centered studies for mobile devices, as well as user evaluations and field trials of new applications and services.