ShoeSense: A New Perspective on Hand Gestures and Wearable Applications

Gilles Bailly[1] **Jörg Müller**[1] **Michael Rohs**[2] **Daniel Wigdor**[3] **Sven Kratz**[2]

[1]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[2]University of Munich, Germany
[3]Department of Computer Science, University of Toronto, Canada
[1]{firstname.lastname}@telekom.de [2]{firstname.lastname}@ifi.lmu.de [3]dwigdor@dgp.toronto.edu

ABSTRACT
When the user is engaged with a real-world task it can be inappropriate or difficult to use a smartphone. To address this concern, we developed *ShoeSense*, a wearable system consisting in part of a shoe-mounted depth sensor pointing upward at the wearer. ShoeSense recognizes relaxed and discreet as well as large and demonstrative hand gestures. In particular, we designed three gesture sets (*Triangle*, *Radial*, and *Finger-Count*) for this setup, which can be performed without visual attention. The advantages of ShoeSense are illustrated in five scenarios: (1) quickly performing frequent operations without reaching for the phone, (2) discreetly performing operations without disturbing others, (3) enhancing operations on mobile devices, (4) supporting accessibility, and (5) artistic performances. We present a proof-of-concept, wearable implementation based on a depth camera and report on a lab study comparing social acceptability, physical and mental demand, and user preference. A second study demonstrates a 94-99% recognition rate of our recognizers.

Author Keywords: Wearable, gestures, gesture set, shoe, sensor placement, mobile.

ACM Classification Keywords: H.5.2 [Information Interfaces And Presentation]: User Interfaces - Interaction styles

General Terms: Design, Experimentation, Human Factors

INTRODUCTION

Using a mobile device when the user is engaged with a real-world task is sometimes inappropriate (meeting, family dinner, church), difficult (walking, running), dangerous (driving) or virtually impossible (the blind) [1]. Wearable technologies offer the promise of an 'always available' computer, but the requirement of a physical mount point for sensors can present challenges for acceptability [11,17]. To address this concern, we developed *ShoeSense*. ShoeSense is a wearable system consisting in part of a shoe-mounted depth sensor pointing upward at the wearer (Figure 1). On-shoe sensor placement has been demonstrated in commercial applications [3], and has the advantage of not being visually apparent, while at the same time affording a view suitable for the detection of hand and arm gestures.

To demonstrate the efficacy of on-shoe placement, we present three novel sets of hand-gestures (*Triangle*, *Radial*, and *Finger-Count*) designed for this setup. These gestures can be performed without visual attention. The characteristics of ShoeSense are illustrated in five scenarios: ShoeSense makes it possible (1) to quickly perform frequent operations without reaching for a handheld (such as answering the phone, reading emails or changing songs); (2) to perform simple operations without disturbing others (such as routing incoming calls to voicemail, activating silence mode, and sending predefined messages); (3) to alleviate some of the limitations of mobile devices (such as the *fat finger problem* [47]); (4) to provide support for accessibility such as home automation; and finally (5) to do artistic performances.

In addition to enhancing social acceptability, ShoeSense introduces a novel and unique perspective (from the shoe) making it possible to recognize discreet and relaxed as well as large and demonstrative gestures without the need for cumbersome hats or body-mounted sensors. The setup provides a large operating volume for performing gestures and does not constrain body movement. In addition to making wearable sensors more socially acceptable, on-shoe mounting offers several benefits. (1) Shoes are 'always available', reducing the need for artificial wardrobe additions such as a pendant or cap. (2) Reduced risk of accidental occlusion by other clothes. (3) Improved image stability. (4) Reduced risk of damage. (5) The weight, rigidity, and large volume of shoes better affords storage of electronics. Finally, (6) the shoe can be used for harvesting power [29].

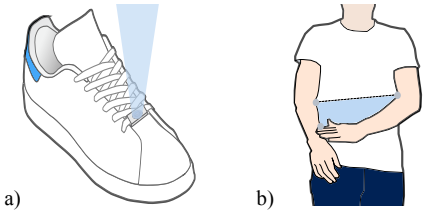

Figure 1:a) ShoeSense perspective. b) Triangle gesture.

In this paper, we present and analyze ShoeSense, including three gesture sets designed for simple eyes-free execution and detection by a shoe-mounted sensor. We describe a proof-of-concept prototype based on a depth-camera to recognize these gestures, present its limitations and demonstrate different classes of applications. We then address the social acceptability of ShoeSense gestures by reporting the results of a lab study comparing the three gesture sets. Finally, we report the results of an experiment validating our recognizer's suitability for a shoe-mounted depth camera.

The primary contributions of this paper are:

- A novel wearable platform enabling a large variety of gesture-based applications.
- A new perspective for eyes-free gestural interaction enabling discreet and relaxed as well as large and demonstrative gestures.
- Three novel gesture sets, designed for a shoe-mounted camera, and experimentally validated as socially acceptable, easy to perform, and robustly detectable.

To situate ShoeSense in the context of existing research, we first review related work.

RELATED WORK

In creating ShoeSense, we sought to build upon the extensive related work in the area of wearable computing. We review two sub-areas of work in that space: the use of body-worn cameras for gaining context, and the use of cameras to enable gestural interaction. We then discuss works, which have inspired the design of our gestures, as well as projects, which have utilized foot-based interaction.

Body-Worn Cameras for Contextual Information

Wearable visual computing uses body-mounted cameras to capture contextual information. In this regard, [36] reports several studies proposing different locations to attach standard video cameras to the wearer's clothing. Locating the camera near the eyes, for example on a hat (WUW [37] or [16,45]), yields images that closely match the user's perspective. Other work has shown merit in placing the camera on users' shoulders [35], on their chest (StartleCam [21], SixthSense [37] or Gesture Pendant [50]), and in multiple other locations on the body [46]. While these placements offer several advantages, each requires the wearer to add elements to their wardrobe or to don technology in visually apparent areas of the body. Our goal in designing ShoeSense was to reduce the social burden of wearable computing by placing the sensors in a visually unobtrusive location, and in an article of clothing typically worn throughout the day. Earlier work has proposed mounting a camera on the user's shoe [15]. Unlike ShoeSense, however, this camera was aimed towards the ground and used to track user activities rather than to enable gestural input.

Gestural Interaction with Body-Mounted Cameras

Several projects have proposed body-mounted cameras to support gestural interactions. SixthSense [37] can project digital information atop physical objects, which can be adjusted using hand gestures captured by a body-worn camera. Gesture Pendant [50], HoverFlow [28] and Imaginary Interfaces [18] also use a chest-mounted camera for capturing hand gestures. All these techniques force the user to perform gestures with the hands at the height of the chest, which can be tiring and draws attention from others. Further, chest-mounting electronics have been shown to introduce serious social issues, suggesting the need for a more discreet placement [11,17].

Starner et al. demonstrated the use of a downward cap-mounted camera for recognizing sign language [49]. WristCam [52] and PinchWatch [33] sense subtle finger gestures with a camera mounted on the wrist [52] or the chest [33], but these cannot capture large and demonstrative gestures. Finally, Ni et al. propose to capture very small gestures in the context of 'disappearing' mobile devices thanks to a one-pixel camera [39]. While relevant, our goal in this work is to minimize the disruption of the introduction of a body-worn camera for 'always available' gesture detection – thus, these projects serve as inspiration for interaction, but not for the design of our hardware and gesture sets which its unique perspective required.

Gesture Design

In designing the gesture language for ShoeSense, we considered several projects that demonstrated advantages of gestural interaction [7, 9, 30, 53]. For instance, Baudel et al. [9] proposed a set of free-hand gestures in Charade for controlling a slide show. Marking menus [30] used an efficient set of 8 radial gestures with a pen or mouse. Finger-Count menus allow the user to perform two-handed and multi-finger interaction on a multi-touch surface [7]. Pinch gestures (for instance in [53]), with one or two hands, have been shown to be easy to perform and to recognize.

In each case, successful gestures have been designed based on consideration of the limitations of sensing, context of use, and use case. Another approach consists of allowing study participants to define the gesture language [44, 54]. While informative, this approach is less suitable for ShoeSense, where gestures must be carefully designed to ensure robust detection given the placement of the camera. Throughout this paper, we use the *Wu* formulation of the phases of a gesture: *registration, continuation*, and *termination* [55].

While several projects have included the use of feet for interaction ([3, 5, 24, 40]), our work uses the foot as a mounting point for a sensor of hand and arm-based gestures. Even though ShoeSense could be enhanced in the future to also detect interactions similar to those described in previous work, our focus is on leveraging the advantages of body-mounted gesture systems while reducing the social burden inherent in wearing digital technology on the body.

Building on this earlier work, we now describe ShoeSense, and the gestures we designed, which take advantage of its unique perspective.

SHOESENSE

ShoeSense consists of an upward-oriented sensor mounted on a shoe. We first discuss the advantages locating the sensor on the shoe. We then describe its I/O capabilities. Finally, we describe the implementation of our device and software.

Shoe-Worn Device

As discussed above, the primary advantage of using shoes as a mounting point is that this setup can enhance the social acceptability of the technology. In addition, mounting a body-worn camera on the shoe offers practical advantages:

Wardrobe integration. Shoes are typically worn continuously throughout the day [23], ensuring that the camera is commonly available. This alleviates the need to artificially add elements to the wardrobe, such as a cap, pendant, or other enclosure.

Reduced occlusion. Sensors worn on the torso face the possibility of occlusion by jackets and other occasional-use clothing. In contrast, a mounted shoe camera has a reliable, unobstructed view of the area in front of the body.

Image stability. Because the foot is planted on the ground, even while walking, this location affords high physical stability of the camera.

Low maintenance. In comparison to t-shirts, pants, or even jackets, shoes are rarely laundered. This is an advantage for ShoeSense as washing machines are a harsh environment for wearable devices [24].

Comfort. The weight of visual wearable devices, even if very light, can be cumbersome for users when they are worn on the top of the body. Further, the added volume and rigidity of shoes provides an opportunity to safely place wearable computing equipment [3].

Discreetness. ShoeSense does not disrupt unrestricted and natural views of the face, which is an important criterion in social interaction [35] and a major difference with devices such as those using head-mounted cameras [35].

Energy. Power distribution is a major challenge of wearable devices [29, 31, 48] as there is a trade-off between capacity and weight/size of batteries. Previous studies [31, 48] have shown that the foot is an excellent location for generating and storing human power when the user is walking [29].

ShoeSense shares some problems of visual wearable devices concerning operation in harsh outdoor conditions. For instance, rain or extremes in lighting can degrade sensor performance. Mud or dust can also affect ShoeSense, as the sensor is located close to the ground. Moreover, the field of view of the camera can be partially occluded by loose clothes or objects (carrying a bag). Additionally, the relative placement of hands in the camera's field of view is more variable than with other arrangements. Finally, the integration in all of the user's shoes could be prohibitive (though hardware constantly gets cheaper). Therefore, in a practical design we envision a small and easily-pluggable device, which can quickly be attached to different shoes.

Gestural Input

ShoeSense is a gesture-based system. The location of the sensor on the shoe offers a novel perspective for detecting and recognizing gestures. A pilot study confirmed that an RGB camera with a traditional field-of view (47°-53°) was sufficient to capture arms' locations within the entirety of their dynamic reach envelope [22]. This is not the case for a camera on the chest, given its close proximity to the arms and hands. ShoeSense thus provides a large detection volume enabling a wide variety of gestures to be observed.

Compared to traditional chest cameras, ShoeSense does not force users to make superfluous arm movements to reach the field of view of the camera, which can be exhausting and attention-grabbing [43]. ShoeSense gestures are more relaxed as they can be performed close to the resting position of the hands, making them both more comfortable and discreet. The large viewing area of a shoe-mounted camera can also enable large and demonstrative gestures, such as may be useful to dancers, stage actors, or musical performers [13]. However, in this article we mainly focus on relaxed and discreet gestures for everyday use. After repeated experimentation, we developed three sets of gestures: *Triangle*, *3D Radial* and *Finger-Count*. Triangle and Finger-Count are examples of two-handed gestures, while 3D Radial is a one-handed gesture.

Triangle Gestures

Triangle gestures are a set of two-armed poses formed by creating a triangle with the arms and torso. Triangle gestures are *registered* by placing a hand atop the opposing arm (Figure 2). In pilot testing, we found that users and system could robustly distinguish between 5 *registration* poses (Figure 2). Once initiated, each of the gestures may be *continued* by modifying the shape of the triangle, by sliding the hand along the arm, or by rotating the 'triangle' in space. The gesture is *terminated* by removing the hand.

The triangle poses are static making it possible to recognize them within only one frame. Moreover, they provide an absolute reference: users can associate specific commands to specific locations on their arm through proprioception. Finally, the inclusion of a continuation phase means the triangle gesture can be used to control a parameter. This continuation can be mapped absolutely, or relatively to allow fine-grained precision with clutching.

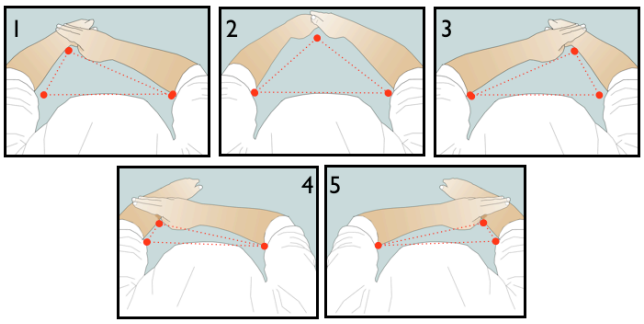

Figure 2: Triangle gesture set. Apex of each triangle is the wrists (1,3), hands (2), and arms (4,5) respectively.

Pinch Registration Pose

A pinch registration pose consists of forming a "hole" in the hand by touching the thumb with the index finger [53]. It is an excellent registration pose for gestural interaction as it is easy to perform, easy to recognize, different from daily life gestures, and does not require a timeout [33, 53]. Pinch can be combined with triangle gestures. For instance, when the user makes a triangle gesture, an audio indication of its function could be given, and the function is only executed once a pinch is performed. The pinch also acts as the registration pose for our remaining 2 gesture sets.

3D Radial Menus

3D radial menus are an extension of 2D radial menus, formed by extrapolating the options of the 2D menu into 3D space [30]. While 2D radial menus can contain up to 8 commands, a 3D radial menu can contain up to 26. To make these gestures unambiguous, the menu is *registered* by a pinch (Figure 3). A command is then executed by *continuing* the gesture in the direction of the desired menu item, and *terminated* by opening the pinch. We argue that this extension has the promise to maintain the properties of marking menus (simple, scale-independent and fast [6, 30]) without degrading memorization [6]. Because 26 options may be sufficient to remove the need for nesting, it is also possible to select and to control a command in the same gesture, as with control menus [42], enabling easy chunking of gestural actions into complex commands [12].

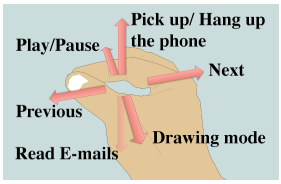

Figure 3: 3D radial gesture set (Left, Right, Front, Back, Up, Down) mapped to frequent actions on a smart phone.

Finger-Count gestures

These static gestures are inspired by [7, 10] and consist in expressing a unary number by extending that number of fingers (**Figure 4**). Considering only the number of fingers rather than their identity makes the technique easier to understand and gives additional physical flexibility [7, 10]. To avoid false positives, we again use the pinch registration pose. Finger-Count gestures are performed on the dominant hand and are interpreted only if a simultaneous pinch is detected on the other hand, or if an appropriate 3D radial menu item is currently held in the other hand. Finger-count can also be used as a modifier for radial menu selections, similar to [32], or for triangle gestures. For instance, a left arm triangle gesture with 5 extended fingers could mean: "Call" (left arm gesture) the contact id "5" (Finger-Count).

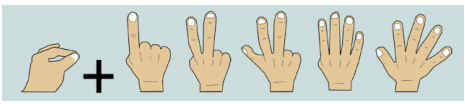

Figure 4: Finger-Count gesture set. Pinch on one hand (registration) and extended fingers on the other hand.

In summary, we propose three gesture sets (*Triangle*, *3D Radial*, *Finger-Count*) as gestural input for ShoeSense and several possible gesture phrases (Triangle + Pinch; Triangle + Finger-Count; Radial + Finger-Count) to select discreet and continuous input. We now consider possible output modalities of ShoeSense.

Output Modalities

Although our proof of concept device is large, we envision that further engineering could enable manufacturing of a device small enough to be smoothly integrated into a shoe, making it difficult to have a traditional visual interface. Interaction with a miniaturized device is then limited by the degree to which user interface affordances can be presented and feedback can be provided. Similar to Ni and Baudisch, we discuss three elementary output modalities attached to the device itself [39]. We then propose two different annexing modalities with higher bandwidth.

Limits of Shoeborne Indicators

As our goal is to locate hardware within the shoe, a speaker is not appropriate as it is far from the ears and has limited privacy. In contrast, tactile (e.g., vibrators) or visual feedback (e.g., LED lights) are suitable for the class of applications where the user is already familiar with the available actions. Tactile is more private (and the foot is fairly sensitive), while visual feedback is less obtrusive. However, each forces users to learn a new language and is low bandwidth. For this reason, shoeborne indicators are only used for confirming that a gesture has been recognized or indicating the current status (i.e. on/off; in progress, etc.).

Opportunistic Audio Channel Annexing

Annexing is the process of serendipitously expanding the I/O capabilities of a system by using nearby I/O components [41]. While loudspeakers are inappropriate, headphone audio can provide a useful feedback channel. This would require an additional device, but would provide greater bandwidth and be easy to interpret. For instance, the system can play the name of the menu items during exploratory gestures. It can also be an implicit part of the application response (e.g., in MP3 players there is no need for a "volume up" indicator, it just gets louder). Commercial applications have already demonstrated wireless connections between shoe-mounted sensors and iPod devices [1]. Annexing of that iPod device's audio channel, where appropriate, could enable this feedback mechanism.

Opportunistic Display Annexing

To enable visual output beyond simple LED indicators, a ShoeSense device could annex mobile device displays, or larger displays in the nearby environment. ShoeSense could then be thought of as a peripheral input device for applications running on those other devices. Users perform mid-air gestures and receive visual feedback on the display. While audio feedback is serial, visual feedback provides random access and a more direct physical mapping. A limitation of annexing a handheld screen is that holding the device limits the users' possible input gestures.

Implementation

To develop a proof-of-concept implementation of ShoeSense (Figure 5), we used a BeagleBoard, a small (82.5x82.5mm), single-board computer with a 1 GHz ARM Cortex-A8 processor and 512 MB onboard memory. We used a depth camera (Microsoft Kinect) as a sensor to enable recognition of hand gestures. The distance between the foot and resting position of the hand (~90cm) is ideal for capturing high quality images given the stock Kinect's lensing. When running on the BeagleBoard, our prototype processes depth images at a rate of about 5 frames per second. This is sufficient to demonstrate the basic interaction techniques during mobile use of the system.

Software

ShoeSense runs Linux, uses OpenNI for capturing depth images and OpenCV for vision. It leverages the relatively fixed distance of the hands/arms from the shoe for simple background subtraction by thresholding the depth image (we use a depth range of 90cm ± 20cm).

Triangle gestures

Triangle gestures are detected if a large inner contour (green color) is found in the image pixels (Figure 6-a). They are then recognized by identifying the left (p_l), right (p_r), and topmost points (p_i) of the inner contour as well as the topmost point (p_o) of the outer contour (Figure 6-a). The orientation and the norm of the vectors [Pi ; Po] and [Pr ; Pi] are sufficient to distinguish the 5 triangle gestures. Finally, we use the ratio of ||PiPl|| / ||PoPl|| to precisely control a value. While the algorithm is quite simple, it is robust as each component is easily detected, with a low error rate.

Pinch gestures and Radial strokes

We recognize pinch gestures by using the algorithm proposed in [53]. It consists of detecting an inner contour (red) in the arm contour. A radial gesture is interpreted by tracking the location of the pinch over time (Figure 6-b).

Finger-Count gestures

As described above, we limit finger-count detection to frames, which also contain a pinch registration pose as shown Figure 6-c. Several methods have been proposed to count fingers [8, 14, 51]. From pilot studies, we chose the k-curvature approach [51] with k=5 because it is simple to implement, does not require RGB images and provides good results even if the hand if slightly tilted.

While relatively simple, this combination of gestural primitives enables a significant set of possible applications.

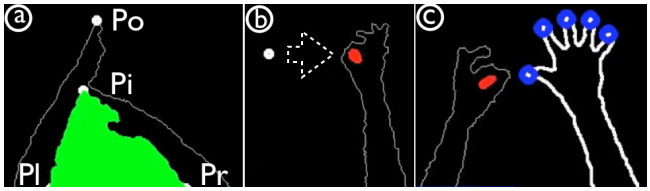

Figure 6: (a) A user performing a hand-to-left-wrist triangle; (b) a 3D radial gesture; (c) a 5 finger-count gesture.

APPLICATIONS

Various interaction scenarios give rise to several classes of interaction. We now review these classes, along with several envisioned sample applications. We also sketch several interaction techniques for mobile devices (Figure 7).

Frequent & Favorite Operations

ShoeSense users can quickly execute favorite/frequent operations such as "answer a phone call", "previous/next track", "volume up/down." Indeed, ShoeSense does not require reaching for the phone in the pocket or bag which requires time and can lead for instance to missed phone calls [4]; to initiate an interaction users simply perform a gesture.

Favorite or frequent operations with mobile devices generally require limited feedback and can involve MP3 player (play/pause, previous/next, repeat one/all, etc.), phone (pick up/hang up a phone call, call favorite contacts, etc.); messages (reading your most recent e-mails or tweets, etc.) or monitoring operations (heart rate, blood pressure, etc.). While several gesture-command mappings are possible, we propose to use Finger-Count to select among function types at a root menu level (e.g., *MP3*, *Phone*, *Messages*, *Monitoring*), which contains 5 or fewer categories. A Radial gesture would then select commands within the selected category (necessary as any given category can contain a larger number of commands). Moreover, as Finger-Count requires a pinch in one hand, it makes it possible to perform these two gestures simultaneously, increasing efficiency and avoiding some possible confusions of mode switching. Finally, Triangle gestures could then be reserved for universal parameters, such as volume, brightness, etc.

Inexact and Inattentive Operations

Hudson et al. describe inexact and inattentive interaction as simple and common operations with a mobile device that can represent a disruption form another activity involving people (meeting, family dinner, etc.) or from the primary task of the user (for example walking while carrying a bag) [26]. ShoeSense has several advantages for controlling functions that fall into one of the following categories.

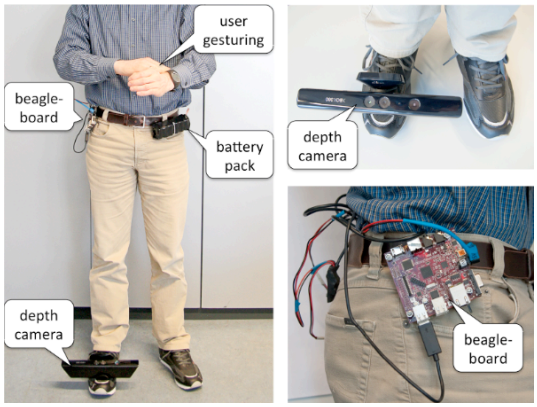

Figure 5. The ShoeSense prototype. The system is wholly contained as a wearable computer.

Fast interaction. As mentioned previously, users do not need to grasp a device. So they can, for instance, quickly silence a ringing cell phone minimizing the time of the disturbance.

Acquisition-free input. The user's hands do not need to hold a special device for interaction [33]. Not only can the user easily return to his or her primary task, but also the hands become free more quickly to react to critical situations.

Eyes-free interaction. Our gesture sets are based on gross motor control and contain gestures that do not require visual assistance for performing them [7, 26, 30]. Users can thus stay concentrated on their primary task. From a technical perspective, this property is useful in wearable scenarios, as screens generally consume a lot of energy.

Operations that would benefit from eyes-free input include "ignore interruptions" or "activate silence mode." A limited number of special operations can be mapped to eyes-free menus. This might include informing favorite recipients (wife, friends) or participants of the next meeting (thus no need to enter the name) of a potential delay.

Interaction Techniques for Mobile Devices

Interaction with mobile devices entails several limitations due to the small size of the screen and the "fat finger problem" [47]. Annexing ShoeSense for off-screen input [41] can enhance the capabilities of the mobile device by separating the input and the output, and also by providing more degrees of freedom.

Separating Input & Output. Inspired by [19, 20], ShoeSense can serve as an input device for a wristwatch (pointing task, command selection and widget control) given a wireless connection between the devices. For instance, as soon as a pinch gesture is detected, the user can control a cursor on the watch by moving the pinched hand in the air. The C/D gain can be adjusted with vertical movement to enhance precision or speed (similar to Figure 7-b). This off-screen interaction technique avoids occlusion on the watch screen. ShoeSense can also be used to select up to 5 different commands (Figure 7-d). Finally, users can control graphical widgets, for instance a slider on a wristwatch (similar to Figure 7-a) by touching the device (left-wrist triangle gesture detection) and moving the finger along the arm (modifying the shape of the triangle). These three interaction techniques are pairwise compatible with one another thanks to the design of the 3 different gesture sets.

Increased Degrees of Freedom. By enabling users to perform mid-air input gestures, ShoeSense provides a novel modality for interacting with mobile devices. For instance, it can help to perform large distance map navigation on a mobile phone. The touch-sensitive surface allows for limited panning of the map until the finger reaches the screen edge. Then, instead of performing inefficient clutching [25], users can continue to pan "in the air" with large horizontal movements and zoom in/out with vertical movements. ShoeSense is useful for navigating in a 3D scene. For instance, (x,y,z)-translation of the camera can be performed with 3D radial gestures (Figure 7-b) and rotation with Triangle gestures (Figure 7-a).

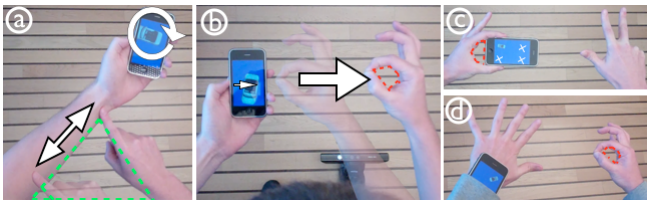

Figure 7: 3D control on mobile devices. (a) Moving the finger along the arm (modifying the shape of the *triangle*) rotates the 3D object. (b) x, y and z translation with *3D radial* gestures. (c) The user chooses a predefined control with a *finger-count* gesture (3 fingers on the right hand + a pinch formed by the left hand and the device). (d) *Finger-count* gestures to select commands on a watch (simulated by an iPhone on the wrist).

Dynamic C/D Gain Control. As previously described, the vertical height of the pinch gesture could adjust the C/D gain of cursor input to a watch device. Conversely, the gain of touch input can be modulated by the surface area of a triangle gesture: To enhance precision while touching their phone, users simply place their other hand on the wrist.

Support for Accessibility

Several projects have been proposed to support accessibility with visual wearable devices [10, 16, 50]. Based on this literature, we imagine the same possible applications for ShoeSense. For instance, ShoeSense could help elderly or disabled people (blind or partially sighted people) to be independent in their home or to be medically monitored [50]. In a multimodal approach, gestures could then be used to delimit the beginning and end of a speech command [10]. Finally, ShoeSense could recognize IR LEDs or 2D tags located on the ceiling, when combined with an IR or RGB camera [16]. Users can then be guided via audio-feedback.

Demonstrative Scenarios

While we focused on applications requiring relaxed and discreet gestures, it is also possible to imagine applications with large and demonstrative gestures such as motion capture (similar to [46]), gesture analysis (golf, baseball), or performance. For instance, air guitar performances [24] or dancing performance (inspired by Adidas Megalizer [1]).

SOCIAL IMPLICATIONS OF SHOESENSE

We distinguish demonstrative scenarios and everyday usage. In the second case, it is essential to consider the social acceptability of the wearable device, which is the primary motivation for the placement of the sensor on the shoe. Commercial products have demonstrated the social acceptability of an appropriately sized, shoe-worn sensing device [3]. Of potential additional concern are the privacy implications of an always-on camera pointing upward from a user's shoe.

The objective of ShoeSense is not to store images (as in [16, 21, 37, 45]), but rather to recognize hand gestures. As such, our prototype uses a depth camera, which has a far

lower resolution than an RGB camera. Moreover, ShoeSense also limits the amplitude to a depth interval of about 40cm, making it virtually impossible to take inappropriate photos to identify people, or the environment. In a commercial realization we envision the complete image acquisition and recognition process to be encapsulated in a single-chip device that only emits the recognition result, but not the raw images. In this way, ShoeSense does not pose a serious privacy risk. Despite this technical possibility, concerns about privacy might still exist. Mann argues that fixed cameras (like surveillance cameras) are generally less acceptable than wearable cameras because users know they are not in private when somebody else is present [34]. The presence of the sensor could be made visually apparent with suitable design elements on the shoe. One way to achieve this is to use a bright LED, which people easily interpret "as an icon for an active camera" [34]. The size of the sensor also has an impact on social acceptance.

USER STUDY: SOCIAL ISSUES & USABILITY

While reducing the obtrusiveness of the sensor may improve social acceptance, there remains the possibility that gesturing in-air is a social concern. We thus designed an experiment intended to elicit participants' reactions to this modality. In addition, we wished to evaluate our gesture sets for their mental and physical demand, as well as user preference. We sought absolute measures of *social acceptability*, *mental* and *physical exertion*. These measurements came in the form of a questionnaire administered after participants had performed all of the gestures. We further recorded users performing these gestures so that we could have a baseline for the development and refinement of our recognizer system.

Questionnaire. The questionnaire (and the design of the experiment) is inspired by Rico et al. [43]. The questionnaire is comprised of the same questions meant to elicit social acceptability. Rico et al. distinguish social acceptability by two factors: *audience* (alone, partner, colleague, friend, family, stranger) and *location* (home, street, driving, passenger, pub, workspace). We omitted driving, because two of our gesture sets require two-handed interaction. Finally, instead of asking participants if "they would be willing to perform the gesture", we used a 10-point Likert scale to refine the results.

To measure physical demand, we used the Annex C of the ISO-9241-9 [27] providing rating scales for finger, wrist, arm, and shoulder effort. To measure mental demand, participants were asked to answer the gestural interaction questions of the NASA-TLX [38]. We also used a 10-point rating scale for these questions to increase their fidelity. Finally, we asked users to order gesture sets by preference.

Gesture. We compared triangle, finger-count and radial gestures. The set of radial menu gestures was limited to 5, in order to match the number of triangle and finger-count gestures, and to avoid biasing *effort* scores. No baseline gesture set could completely cover the features of the proposed approach (always available, hands-free operation, no need to hold a phone, discreet). As we tried to build on proven components such as the common pinch and finger-count gestures, this provides a sort of baseline.

Procedure. To evaluate the social acceptability of the three gesture sets, the participants watched a video showing a standing actor performing each gesture in front of a white wall [43]. Then, participants were asked to perform the gesture 3 times. The experimenter observed to ensure that the gesture was performed correctly. After the completion of all gestures, the questionnaire was administered.

Participants. 12 participants (7 female), aged from 21-37 (m=28, sd=5.3), were recruited from the local community. Except two users having already played with the Xbox Kinect, none had prior experience with in-air gestural interfaces. Each received compensation of €10.

Apparatus. A Kinect connecting to a PC was mounted onto the shoe of each participant. Instead of using it to recognize gestures, it was used to capture depth videos used later to develop a recognizer for these "naturally" performed gestures. Indeed, using a specific recognizer could have had a strong impact on how users perform gestures and so also on their ratings of those gestures on each of our scales. For this reason, this study is independent of a recognizer. The follow-up study focuses on accuracy after having improved our recognizer according to our observations.

Setting. Our prototype was too large to avoid confusing "social comfort" when measured in public. For this reason, we decided to perform a lab study, focusing on the social acceptability of gestures sets rather than the prototype. A field study should be run in order to evaluate the social acceptability of gestures and the envisioned system in public.

Design. We used a within-participant design: the order of the gesture sets was counter-balanced across participants using a Latin-square. The design can be summarized as:

12 participants x 3 gesture sets x 5 gestures per set x

3 repetitions per gesture = 540 total gestures performed.

Results

We used the Kruskal-Wallis (KW) test to analyze the non-parametric data collected for *social acceptability*.

Context & gesture set. The mean acceptance ratings across all contexts is above 5 for each gesture set (m_{Radial}=7.8; $m_{Finger\text{-}Count}$=7.2; $m_{Triangle}$=5.9). The Kruskal-Wallis (KW) test reveals a significant effect (χ^2=15.3, $p<0.01$) on social acceptability for gesture sets. Pair-wise comparisons show that Radial is significantly more acceptable than Triangle. Radial and Finger-Count gestures are rated as likely to be used (>5) in all contexts. Triangle is likely to be used in *home*, *pub* and *workspace*. However, a deeper analysis reveals that 50% (6/12) of participants would be willing to perform Triangle gestures in *street* or *transport*. These results are illustrated in **Figure 8**.

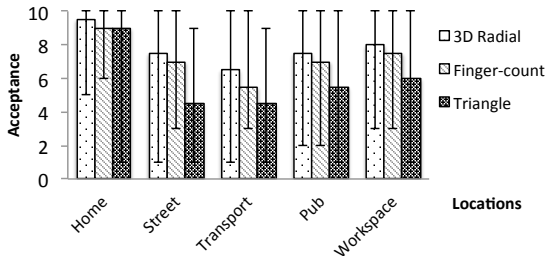

Figure 8: Social acceptability by location and gesture set (10-point scale, 10 is the best). 95% confidence interval marked.

Audience and social acceptability. The mean *social acceptability* score across participants indicates all gesture sets are acceptable in front of all audiences (mean >6) except Triangle gestures, in front of strangers (7 of 12 rated < 5). These results are illustrated Figure 9.

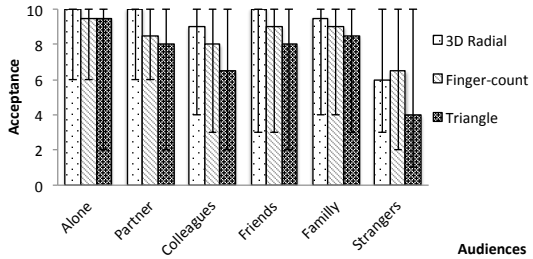

Figure 9: Social acceptability by audience and gesture set (10-point scale, 10 is the best). 95% confidence interval marked.

It is interesting to note that the standard deviation of the acceptability ratings of the triangle gestures is much larger than for 3D radial gestures or Finger-Count. This is true for both location (location: sd_{Radial}=2.1; $sd_{Finger\text{-}count}$=1.9; $sd_{triangle}$=2.8) and audience (audience: sd_{Radial}=1.6; $sd_{Finger\text{-}Count}$=1.6; $sd_{Triangle}$=2.6) and indicates that there was less consensus among participants about the acceptability of the triangle gestures.

	Finger	Wrist	Arm	Shoulder
Radial	2.5	2.0	3.0	2.5
Finger-Count	3.0	1.5	2.5	2.5
Triangle	1.0	1.0	3.0	3.5

Table 1:Physical demand (10-point scale, 1 is the best).

Physical demand. Physical demand was rated for each gesture set on various body parts using Annex C of the ISO-9241-9 [27]. The average physical demand is low for each gesture set (m_{Radial}=2.5; $m_{Finger\text{-}Count}$=2.4; $m_{Triangle}$=2.1). Details for fingers, wrist, arm and shoulder are reported in Table 1. The KW test ($\chi^2 = 9.3$, $p < 0.01$) followed by a pairwise comparison shows that Finger-Count (3) is more exhausting for fingers than Triangle (1). Three of 12 participants indicated that extending 3 fingers was a little bit more "tricky". All other ratings are not significantly different from one another at the p=.05 level.

Mental demand. Mental demand was measured using the NASA-TLX [38] methodology. The average mental demand is also quite low for each gesture set (m_{Radial}=2.0; $m_{Finger\text{-}Count}$=4.5; $m_{Triangle}$=3.0). KW test reveals no significant difference of mental demand across gesture sets. 7 participants indicated that two-handed gestures (Finger-Count and Triangle) require more coordination. Two participants also indicated that extending 3 fingers requires more mental demand than other Finger-Count postures.

Preference. We asked users to rank-order gesture sets by preference. 6 participants chose Radial as their favorite gesture set and 6 chose Triangle, 0 chose Finger-Count. 4 participants chose Radial as the least favorite gesture set, 3 chose Triangle and 5 chose Finger-Count.

Discussion

Social acceptability. The mean value of *social acceptability* for 3D Radial and Finger-Count is above 7 for all *locations* (except public transport) and *audiences* (except strangers). Triangle has a low acceptance rate (≤6) for the *locations street*, *public transport,* and for the *audience strangers*. The high standard deviation of social acceptance for Triangle is interesting. On the one hand, 2 participants rated Triangle very highly for most *audiences* and *locations*, sometimes higher than gestures of other types, and gave feedback that Triangle gestures are "very natural", "it is like if I look at my watch" (adjusting a watch forms a triangle gesture). On the other hand, 2 participants found these gestures highly unacceptable: "they require too much space for performing gestures especially in the subway". We observed in the videos that some participants performed Triangle gestures too high (arms were almost horizontal), requiring more space and more effort than needed. They probably under-estimated the capacity of the system to capture casual Triangle gestures. This could partially explain this rating. More generally, we were surprised about the differences between participants in rating social acceptability and the reasons given. For instance, 4 participants found it acceptable (≥ 8) to perform gestures in the workplace because "I know my colleagues" while 2 participants found them not acceptable (≤ 3) because "workplace is serious."

Physical and mental demand. The three gesture sets do not require high physical (≤ 3) or mental (< 5) demand. Only Finger-Count, and especially the 3-finger posture, is a little bit more mentally and physically demanding. Finally, these results are more promising than might be expected for Triangle gestures (effort: 2.1; mental demand: 3.0) as they involved arm movements.

Preference. Participants do not share the same opinion about gestures but they seem to agree that Finger-Count is less preferable than Radial and Triangle. Discussions with participants suggest that they do not like to perform gestures which require distinguishing between left and right hands, nor do they like "to think about two different gesture types: Pinch and Finger-Count". More fundamentally, we think participants did not understand the need for a pinch gesture on the left hand (gesture delimitation, not explained to the participants) and thought this artificially increased the mental demand. Pilot studies we performed in the initial steps of this project suggested that Finger-Counting gestures

without pinching were appreciated by participants (but risked accidental command invocation). We plan to investigate Finger-Count gestures without the pinch delimiter, by taking into account the position and orientation of the hand to avoid accidental activations.

Observations for Recognizer Optimization

After the study, we examined the depth recordings in order to improve our recognition algorithms. These observations aided in the development of the algorithms we described earlier in this paper:

Pinch and Radial. The pinch "hole" was sometimes quite small and not always visible due to the low resolution of the Kinect. So a pinch is now terminated *only* when it is not detected during 3 consecutive frames to avoid false recognition during fast motions. This introduces a lag of up to 400ms, which would be reduced by increasing frame rate.

Finger-Count. Again, the limited resolution of the Kinect device can sometimes make finger distinction difficult especially for distinguish four and five fingers due to the proximity of fingers. Future implementations would be well served to use a higher resolution sensor.

Triangle. Our first implementation only considered the triangle shape (Figure 6-a). However, it appeared sometimes difficult to distinguish between arm and wrist triangles. To more reliably distinguish between these two postures we included end of the reference arm (Po in Figure 6-a). This also helps in more precisely controlling a continuous value.

USER STUDY: ACCURACY

The previous study investigated usability of our gesture sets without the constraints of using a specific recognizer. We now aim to validate our gesture recognizer and to determine a baseline gesture recognition rate, both with and without visual feedback, that other researchers may improve upon.

Design. 12 novel participants (aged from 25-32 were asked to perform each gesture 5 times for each condition: *with* and *without* visual feedback. In both conditions, they were informed if the gesture was correctly recognized or not, but only in the visual feedback condition, participants saw the performed gesture on a screen (which was removed for second condition). A Kinect was installed over the participant's shoe and connected to a PC to enable a higher frame rate. The system recognized a vocabulary of 15 commands (5 gestures in each set). All participants had 2 minutes of training for each set and started with the visual feedback condition first (to act as training for the eyes-free condition). The stimulus consisted of the name of the gesture. The order of sets was counter-balanced between participants.

Results. ANOVA reveals a significant effect for gesture sets ($F_{2,22}=5.62$, $p<.001$). A post-hoc Tukey test shows that Radial (99.0%) is significantly more accurate than Triangle (95.0%) and Finger-Count (94.0%). ANOVA reveals no effect on accuracy for visual feedback (with: 95.5%; without: 96.5%).

Discussion

Results show a high level of accuracy (>94%) for the three gesture sets. Moreover, results indicate that our gesture sets allow for eyes-free interaction. However, we were surprised about the difference with the visual feedback condition (with: 95.5%; without: 96.5%). The short training phase (2 min), the order of conditions and the fact that – based on our observations – the participants tried to be more accurate in the second condition (without visual feedback) can explain this unexpected result. As in the previous study, we observed participants sometimes elect to perform less casual gestures in favor of accuracy, especially for Triangle. However, recognition performance is bounded by the resolution of the camera and its frame rate. As these improve, so too will recognition, as will the subtlety with which users can perform their gestures.

MOBILITY AND FUTURE WORK

In this article, we mainly focused on gestural interaction when standing. While seated, finger-count and radial gestures can be recognized and comfortably performed on the left/right side of the legs. The triangle gesture cannot be performed so comfortably. We plan to investigate the ability of ShoeSense to operate in walking or running scenarios. Fitzpatrick & Kemp [15] demonstrate that (1) while walking, one foot is always stable on the ground (swing phase) (2) it is possible to detect these periods of stability with a camera by calculating the spatial derivative, and (3) they can get high quality images for recognizing obstacles and floors. Adding accelerometers and gyroscopes to the prototype (as well as buttons under the shoe) could also help to precisely detect and measure periods of stability.

CONCLUSION

We proposed ShoeSense, a wearable system consisting of a shoe-mounted sensor that enables gestural input. ShoeSense provides an unobtrusive always-available input mechanism that does not constrain body movement. We presented three gesture sets designed for eyes-free interaction. We then demonstrated that these gestures can be freely combined. Our proof of concept implementation shows that it is easy to develop robust recognizers for these gestures. The approach enables a wide range of scenarios, especially by annexing visual wearable devices or by enhancing operations on mobile devices.

We reported the results of two user studies. The first one shows promising results in terms of social acceptability, physical and mental demand, and users' preference of the three gesture sets. It also reveals a strong variability concerning social acceptability and preferences of gestures between participants. The second study shows that the reference implementation is robust and that the recognition rate is between 94-99%, even for eyes-free operation. These two studies led us to conclude that ShoeSense is a viable option for future wearable technology and gestural

interaction. The next step would be confirm these results by performing a field study to evaluate ShoeSense in the "Wild". In particular, to investigate the ability of ShoeSense to operate in walking or running scenarios.

ACKNOWLEGMENT

This work was supported by the Alexander von Humboldt Foundation. We thank D. Guse and R. Walter for the implementation, A. Roudaut, T. Pietrzak, S. Malacria for their useful comments.

REFERENCES

1. Adidas Megalizer: http://popsop.com/44586
2. Amft O. and Lukowicz, P. 2009. From Backpacks to Smartphones: Past, Present, and Future of Wearable Computers. *IEEE Pervasive Computing* 8, 3 (July 2009), 8-13.
3. Apple Inc., Take Nike + iPod on your Run. In *Apple.* As of 2011.9.9 from http://www.apple.com/ipod/nike/run.html
4. Ashbrook, D. L., Clawson, J., Lyons, K., Patel, N., Starner, T. 2008. Quickdraw: the impact of mobility and on-body placement on device access time. *CHI '08*, 219-222.
5. Augsten, T., Kaefer, K., Meusel, R., Fetzer, C. Kanitz, D. .Stoff, T., Becker, T., Holz, C., Baudisch, P. 2010. Multitoe: high-precision interaction with back-projected floors based on high-resolution multi-touch input. *ACM UIST'10*, 209-218.
6. Bailly, G. Lecolinet, E, Nigay, L.. 2008. Flower menus: a new type of marking menu with large menu breadth, within groups and efficient expert mode memorization. *AVI'08*, 15-22.
7. Bailly, G., Lecolinet, E., Guiard, Y. 2010. Finger-count & radial-stroke shortcuts: 2 techniques for augmenting linear menus on multi-touch surfaces. ACM CHI '10, 591-594.
8. Bailly, G., Walter, R., Müller, J., Ning, T., Lecolinet, E. 2011. Comparing Free Hand Menu Techniques for Distant Displays using Linear, Marking and Finger-Count Menus. IFIP *INTERACT'11*. 248-262.
9. Baudel T., Beaudouin-Lafon, M. 1993. Charade: remote control of objects using free-hand gestures. *Commun. ACM* 36, 7, 28-35.
10. Benko, H. and Wilson, A. D. (2010). Multi-Point Interactions with Immersive Omnidirectional Visualizations in a Dome. *ACM ITS '10*. 19-28.
11. Brett, B., MIT student arrested at Logan in bomb scare. *The Boston Globe*. 2011.9.9 from http://www.boston.com/news/globe/city_region/breaking_news/2007/09/mit_student_arr.html.
12. Buxton, W. 1986. Chunking and phrasing and the design of human-computer dialogues, *IFIP WCG*, 475-480.
13. Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R. Volpe, G. 2004. Multimodal analysis of expressive gesture in music and dance performances. *Gest.-based. Com. in Hum. -Comp. Int.* vol. 2915. 357-358
14. Conseil, S. Bourennane, S., Martin, L. 2007 Comparison of Fourier Descriptors and Hu Moments for Hand Posture Recognition. *EUSIPCO'07.*
15. Fitzpatrick, P. and Kemp C. C. 2003. Shoes as a Platform for Vision. *IEE ISWC '03*. 231-233.
16. Foxlin, E. M. 2002. Generalized architecture for simultaneous localization, auto-calibration, & map-building. *IEEE/RJS.* Vol.1. 527-533.
17. Guernsey, L., At Airport Gate, a Cyborg Unplugged. In *The New York Times*. As of 2011.9.9, from http://www.nytimes.com /2002/03/14/technology/circuits/14MANN.html.
18. Gustafson, S., Bierwirth, D. Baudisch, P. 2010. Imaginary interfaces: spatial interaction with empty hands and without visual feedback. ACM *UIST'10*, 3-12.
19. Harrison, C. and Hudson, S.E. 2009 Abracadabra: wireless, high-precision, and unpowered finger input for very small mobile devices. *ACM UIST'09*, 121–124.
20. Harrison, C. Tan, D., Morris, D. 2010. Skinput: appropriating the body as an input surface. *ACM CHI '10*, 453-462.
21. Healey, J., Picard, R. 1998. Startlecam: A cybernetic wearable camera, *Tech. Rep*. 468, October 1998.
22. Hedge, A., Anthropometry and Workspace Design, in DEA 325/651. 2002, Cornell.
23. Hendry J, 1984. Shoes: the early learning of an important distinction in Japanese society, *Europe Interprets Japan*, 215-222
24. Higuchi, H. and Nojima, T. 2010 Shoe-shaped i/o interface. *ACM UIST '10*. 423-424.
25. Hinckley, K. 2007. Input technologies and techniques. Chapter 9. In The *human-computer interaction handbook*, 2nd edition.
26. Hudson S., Harrison, C, Harrison, B., LaMarca, A. 2010. Whack gestures: inexact and inattentive interaction with mobile devices. *ACM TEI '10*. 109-112.
27. ISO/DIS 9241-9. 2000. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 9. ISO.
28. Kratz, S. Rohs, M. 2009. Hoverflow: exploring around-device interaction with IR distance sensors. *ACM MobileHCI'09*. 1-4.
29. Krupenkin, T., Taylor, J. 2011. Reverse electrowetting as a new approach to high-power energy harvesting. *Nat. Commun*. 2:448
30. Kurtenbach, G., Buxton, W. 1991. Issues in combining marking and direct manipulation techniques. *ACM UIST'91*, 137-144.
31. Kymissis, J., Kendall, C., Paradiso, J., Gershenfeld, N. 1998. Parasitic Power Harvesting in Shoes. *ISWC'98*, 132-139.
32. Lepinski, G. J., Grossman, T., Fitzmaurice, G. 2010. The design and evaluation of multitouch marking menus. *ACM CHI'10*. 2233-2242.
33. Loclair, C., Gustafson, S., Baudisch, P. 2010. PinchWatch: A Wearable Device for One-Handed Microinteractions, *ACM MobileHCI'10* workshop on Ensembles of on-body devices.
34. Mann, S. 1995. Privacy Issues of Wearable Cameras Versus Surveillance Cameras. *Newsweek*, 86(11), 21-22.
35. Mayol, W., Tordoff, B., Murray, D. 2000. Towards wearable active vision platforms. *Trans. Syst. Man. Cyber*. V. 3, (Oct. 2000), 1627-1632.
36. Mayol, W., Tordoff, B., Murray, D. 2009. On the choice and placement of wearable vision sensors. *Trans. Sys. Man Cyber.* 39, 2. 414-425.
37. Mistry, P., Maes, P., Chang, L. 2009. WUW-wear Ur world: a wearable gestural interface. *ACM CHI EA* '09. 4111-4116.
38. *NASA TLX*, NASA Ames Research Center, Moffet Field, California, 1988.
39. Ni, T., Baudisch, P. 2009. Disappearing mobile devices. *ACM UIST '09*. 101-110
40. Paradiso, J. and Hu, E. 1997. Expressive Footwear for Computer-Augmented Dance Performance. IEEE ISWC '97, 13-14.
41. Pierce, J., Mahaney, H. 2004 Opportunistic Annexing for Handheld Devices: Opportunities and Challenges. *HCIC04*.
42. Pook, S., Lecolinet, E., Vaysseix, G., Barrilot, E. 2000. Control menus: execution and control in a single interactor. *ACM CHI EA'00*, 263-264.
43. Rico, J. and Brewster S. 2010. Usable gestures for mobile interfaces: evaluating social acceptability. *ACM CHI '10*, 887-896.
44. Ruiz, J., Li, Y., Lank, E. 2011. User-defined motion gestures for mobile interaction. *ACM CHI '11*, 197-206.
45. Schiele B. and Pentland, A. 1999. Attentional objects for visual context understanding. Tech. Rep. 500, MIT media Lab, 1999.
46. Shiratori, T., Park, H. S., Sigal, L., Sheikh, Y., Hodgins, J. K. 2011. Motion Capture from-Mounted Cameras. *SIGGRAPH*. 10 pages.
47. Siek, Roger, Y., Connelly, K. 2005. Fat finger worries: How older and younger users physically interact with PDAs. *INTERACT '05*. 267-280.
48. Starner, T. 1996. Human-powered wearable computing. *IBM Syst. J.* 35, 3-4 (September 1996), 618-629.
49. Starner, T., Wearer, J., Pentland, A. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Trans. Pat. Anal. Mach. Intell*. 20, 12 (Dec. 98), 1371-1375.
50. Starner, T., Auxier, J., Ashbrook, D., Gandy, M.2000. The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. *IEEE ISWC '00*, 87-94.
51. Trigo, T. Pellegrino, S. 2010 An Analysis of Features for Hand-Gesture Classification. *IWSSIP'10.* 412-415
52. Vardy, A., Robinson, J., Cheng, L. 1999. The WristCam as Input Device. *IEEE ISWC* '99, 199-202.
53. Wilson, A. D. 2006. Robust computer vision-based detection of pinching for one and two-handed gesture input. ACM *UIST'06*, 255-258.
54. Wobbrock, J., Morris, M., Wilson, A. 2009. User-defined gestures for surface computing. *ACM CHI '09*, 1083-1092.
55. Wu, M., Shen, C., Ryall, K., Forlines, C., Balakrishnan, R. 2006. Gesture registration, relaxation, and reuse for multi-point direct-touch surfaces. IEEE *TableTop'06.* p. 183.190.

The columns on the last page should be of approximately equal length.
Remove these two lines from your final version.