zSense: Enabling Shallow Depth Gesture Recognition for Greater Input Expressivity on Smart Wearables

Anusha Withana, Roshan Peiris, Nipuna Samarasekara, Suranga Nanayakkara
Augmented Senses Group, International Design Center,
Singapore University of Technology and Design, Singapore
{anusha | roshan_peiris | nipuna | suranga}@sutd.edu.sg

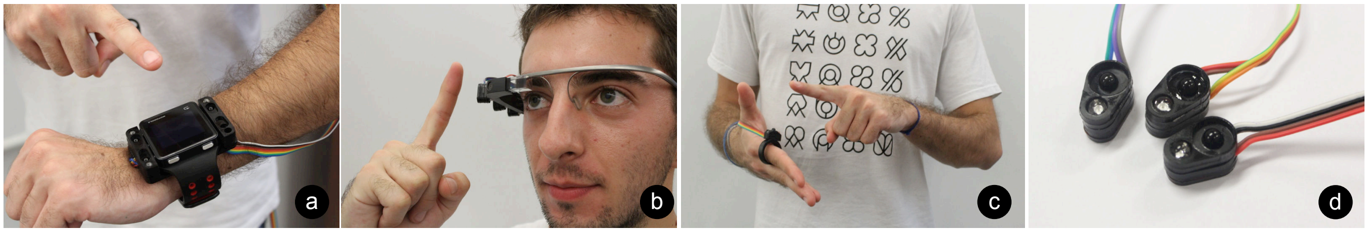


Figure 1: Shallow depth gesture recognition with *zSense*: a), b) Extending interaction space of a smartwatch and a smartglass, c) Enabling gesture interaction on a smart ring, d) *zSense* Sensor-emitter modules.

ABSTRACT

In this paper we present *zSense*, which provides greater input expressivity for spatially limited devices such as smart wearables through a shallow depth gesture recognition system using non-focused infrared sensors. To achieve this, we introduce a novel Non-linear Spatial Sampling (NSS) technique that significantly cuts down the number of required infrared sensors and emitters. These can be arranged in many different configurations; for example, number of sensor emitter units can be as minimal as one sensor and two emitters. We implemented different configurations of *zSense* on smart wearables such as smartwatches, smartglasses and smart rings. These configurations naturally fit into the flat or curved surfaces of such devices, providing a wide scope of *zSense* enabled application scenarios. Our evaluations reported over 94.8% gesture recognition accuracy across all configurations.

Author Keywords

Shallow Depth Gesture Recognition; Interacting with small devices; Smart Wearables; Compressive Sensing

ACM Classification Keywords

H.5.2 User Interfaces: Input devices and strategies

INTRODUCTION

The recent unveiling of smart wearable technologies has led to an increased demand for devices such as smartwatches[1], smartglasses[2], smart rings [24] and other body-worn accessories [25]. These compact devices often have to compromise between reduction of form factor and sophistication of interactions [30]. Among potential solutions, mid air gestures have been one of the leading frontiers [19, 10, 27] as it extends the interaction space beyond the device's surface area. However, many existing approaches [22] require significant processing power, physical space and energy resources that are extremely limited in compact devices. In addition, for smaller devices, shallow depth or close proximity gestures are more desirable [6]. As such, *zSense* focuses on providing greater input expressivity on these spatially limited devices.

In *zSense*, we introduce a non-linear spatial sampling (NSS) technique for shallow depth gesture recognition based on the compressive sensing principle [3, 21] and spatial light modulation [9]. The *zSense* system consists of infrared (IR) sensors and emitters that can be arranged in many different configurations (Figure 2). For example, the number of these units can be as minimal as one sensor and two emitters. NSS technique enables us to overcome the challenges faced by gesture recognition systems for small form factor devices due to following key features:

> *Spatially efficient non-focused gesture sensing:* Typical non-focused gesture sensing requires a relatively larger separation between sensors [8] compared to *zSense*. In the current implementation, five expressive gestures could be detected from a single sensor and two emitters located within a $15mm$ space.

> *Low energy consumption:* In *zSense*, the effective ON time of emitters are minimal due to a modulated spatial lighting pattern. For example, in the current implementation, emitters are on for only 8% of the time. This significantly

[1] `http://goo.gl/G7USq6`, `http://goo.gl/olfxj9`

[2] `http://goo.gl/2nZVEh`

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.
Copyright © 2015 ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702371

reduces the energy consumption compared to existing systems.

Low processing power: The effective data throughput of *zSense* is $1.2kBps$ (with three sensors) or $0.2kBps$ (with one sensor), which generates significantly less amount of data that is needed to be further processed.

Low cost: System uses low-cost off-the-shelf components and standard microprocessors without requiring any special hardware.

With these capabilities, *zSense* enables a broad application space leveraging on limited space and energy available in small scale smart devices. These devices and their unique form factors support different configurations of *zSense* to recognise expressive gestures with high accuracy. For example, embedding only two emitters and one sensor on the curved surface of a ring (Figure 1c) enables recognition of five gestures with over 91% accuracy. This, for example, could be paired with a smartglass to play a game on the go. Three sensors and emitters can be embedded on the narrow space between the screen and the edge of a smartwatch (Figure 1a) to recognize seven gestures with over 91% accuracy. Furthermore, the elongated surface on the temple of a smartglass can be embedded with three sensors and three emitters to recognise eight gestures with over 90% accuracy.

In the next few sections, we introduce the concept and theoretical background of *zSense*; different sensor-emitter configurations; details of the prototype implementation; user evaluation; and application possibilities. The contributions of this paper can be summarized as,

- Introduction of a non-linear spatial sampling (NSS) approach for shallow depth gesture recognition and development of a theoretical framework for *zSense*
- Exploration of various spatial arrangements of *zSense* configurations and identifiable gestures
- Prototype implementation of *zSense*, verification of its performance through user studies and a discussion of application space.

RELATED WORK

Extending the interaction space on smart devices has been a widely researched topic. Computer vision technologies such as 2D cameras [10], markers [4, 26] and commercial depth cameras[3] have been frequently used in the literature due to its ability to track gestures relatively robustly in real time [2]. Harrison et al., in OmniTouch [12] presents a wearable sensor-projection system to enable interaction on everyday surfaces. Here the authors use a depth camera attached to the shoulder to identify various gestures or surfaces for interaction. In [17], the authors use similar depth camera tracking system to provide around the device interaction to investigate free-space interactions for multi scale navigation with mobile devices. However, such computer vision based approaches generally operate by capturing the whole interaction space frame by frame, and processing each frame to identify gestures. As such, these technologies require high computational processing power, high energy for operation and generally a large setup which makes these technologies less desirable for spatially constrained application domains. Furthermore, since camera based systems have a minimum focusing distance, near-device (shallow depth) gestures are not recognisable.

Magnetic field sensing is another frequent technique that has been used to extend the interaction space around mobile devices [18, 1, 7]. Works such as [13] and [16] use external permanent magnets to extend the interaction space around the mobile device. Both works use the inbuilt magnetometers of the device to detect the magnetic field changes around the device as input to the system. Here, [13] uses a magnet on the finger for gestural input and [16] introduces new unpowered devices for interaction. However, the magnetic sensing approach requires instrumenting the user, generally having the user wear a magnet on the finger tip.

Almost all mobile devices are equipped with sound sensors. As such, TapSense [14] is a technology that uses the embedded sound sensor to classify various sounds. It allows the user to perform different interactions by interacting with an object using the fingernail, knuckle, tip, etc. Skinput [15] on the other hand uses the human body for acoustic transmission. Here, the authors introduce a sensor embedded armband to identify and localise vibrations caused by taps on the body as a form of input.

Infrared is another common technology that has been used to extend the interaction space with mobile devices. In SideSight, authors use arrays of infrared sensors attached on the two sides of a mobile device to provide multi-“touch” interaction when placed on a flat surface [5]. Similarly, in [23], authors use infrared beams reflected from the back of your hand to extend interactions with a smart wristwatch. Here, the users interact with the wristwatch by using the back of your hand as an input touch pad.

Additionally, infrared technologies have often been used for mid air gesture sensing as well. In Mime [8], authors introduce a 3D gesture sensor that uses a three pixel infrared time of flight module combined with a RGB camera. Here, the authors introduce tracking 3D hand gestures with the time of flight module, that is combined with the images from the RGB camera to perform finer shape based gestural interactions. However, in Gesture Watch [20], the authors use infrared proximity sensors on a wrist worn device combined with hidden Markov models to recognise gestures for interaction with other devices. The use of infrared in this context requires to light up the whole interaction space [5, 20] which generally requires relatively high energy for the scope of spatially constrained applications. In addition, capturing and processing the whole interaction space would require relatively high computation power [20]. Similarly, Mime [8] operates at a high sampling rate which increases the processing power required for the system.

Our main goal with *zSense* is to develop a low power, low processing overhead and a spatially efficient gesture sensing

[3] `http://goo.gl/wjFcjn`

technology for wearable devices. As such, even though technologies such as depth cameras provide extremely accurate sensing, such technologies require a vast amount of processing power due to the amount of information captured by the camera. Some other technologies such as magnetic sensing requires instrumenting the user's fingers or sound sensing requires a high processing power. In contrast, in *zSense*, we propose NSS to spatially modulate the sensing environment such that a minimal and only required amount of information can be captured for sensing. As such, NSS minimises the required number of sensors and emitters for gesture sensing on devices with limited space. This significantly reduces its power and processing requirements and improves its spatial efficiency.

ZSENSE CONCEPT

zSense concept is based on optimising space, power and cost while being able to recognize a reasonable number of expressive gestures. To this end, we adapted compressive sensing and spatial light modulation to propose the concept of Nonlinear Spatial Sampling (NSS) for gesture recognition.

Non-linear Spatial Sampling (NSS)

Computer vision based gesture recognition systems require a high sensor density since each spatial location is mapped to a pixel sensor (linear sampling). Similarly, other approaches such as IR based systems, equally sample the space with a high number of sensors (high density) [5, 20]. In *zSense*, we use a significantly less amount of emitters with relative spatial displacements (linear or angular) between each other, and temporally modulate the emitters with different lighting patterns. Each sensor records separate measurements of reflected light from a target per each modulating pattern. Since sensor values represents a cumulative sum of the reflected light from a spatial light pattern, recorded data carries spatial information of the target. Therefore this non-linear spatial sampling (NSS) scheme enables *zSense* to minimize the number of sensors and the number of emitters needed, leading to reduced power and signal processing requirements.

zSense Configurations

Any *zSense* system must have at least two IR emitters coupled with one or more sensors. The spatial arrangement of sensors and emitters is a key factor in determining the modulated spatial illumination pattern and the quality of the captured signal. Therefore, selecting an appropriate spatial configuration is vital to *zSense*'s operation. In addition, the accuracy and the number of recognizable gestures increases with the number of sensors and emitters. For smaller devices, a trade off is required for the accuracy and the number of desirable gestures at the design stage.

In *zSense*, we use spatial configurations of emitters to implement NSS. This can be achieved by a linear displacement or an angular displacement. Linear configurations can be achieved in two axes (displacement along front facing axis of emitters will not change the spatial pattern), and angular as the relative angle. In this paper, we describe three configurations to demonstrate these three variations using a linear displacement along one axis (Figure 2a), linear displacements along two axes (Figure 2b) and angular variations (Figure 2c). These configurations naturally support flat and curved surfaces.

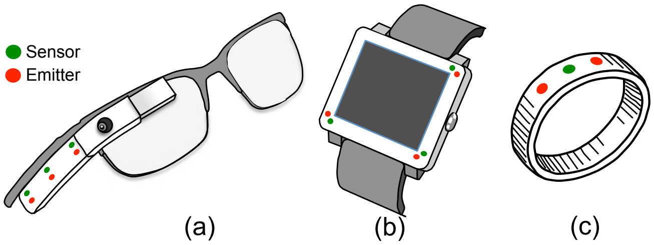


Figure 2: Selected set of spatial configurations: a) One axis linear displacement (3S3E linear), b) Two axes linear displacement (3S3E triangular), c) Angular displacement (1S2E angular).

One Axis Linear Displacement (3S3E Linear): This configuration is studied using three sensors and three emitters arrangement as shown in Figure 2a. This arrangement makes sense for devices with an elongated flat surface such as temple of a smartglass or a tie-pin.

Two Axes Linear Displacement (3S3E Triangular): This configuration is shown in Figure 2b where sensor units are arranged in a triangular shape. This configuration supports devices with a flat surface (such as a smartwatch frame) to perform simple gestures along either axis of the device.

Angular Displacement (1S2E Angular): This is ideal when the space available for interaction has a curved surface, such as buttons, earrings, finger rings etc. For example, Figure 2c shows a configuration where two emitters and a sensor are mounted on a ring. NSS is achieved by leveraging the form of the ring where curvature of the surface introduces an angular displacement between the sensor and emitters in opposite directions.

Gesture Space

We use *Form* based categorisation [29] to name the surface gestures. *Form* describes the gestures that are differentiated with pose and motion and is further divided into six categories, in which we have selected *static pose*, *dynamic pose* and *static pose and path*. These gestures are shown in Figure 3.

Static pose paths shown in Figure 3c can be performed in symmetrical mirror images in case of the sensors are place symmetrically, for an example, swipe far (Figure 3G5) and swipe close (Figure 3G6) can be performed from right to left and left to right. In the same manner, diagonal swipe (Figure 3G7) can be from left to right and right to left. Circular gesture (Figure 3G8) can be clockwise or anti-clockwise.

Gesture Set

Within the gesture space described above, we select a set of gestures (Table 1) based on the following criteria.

Applicability: We include most commonly used gestures such as swipes to be studied across all the considered configurations.

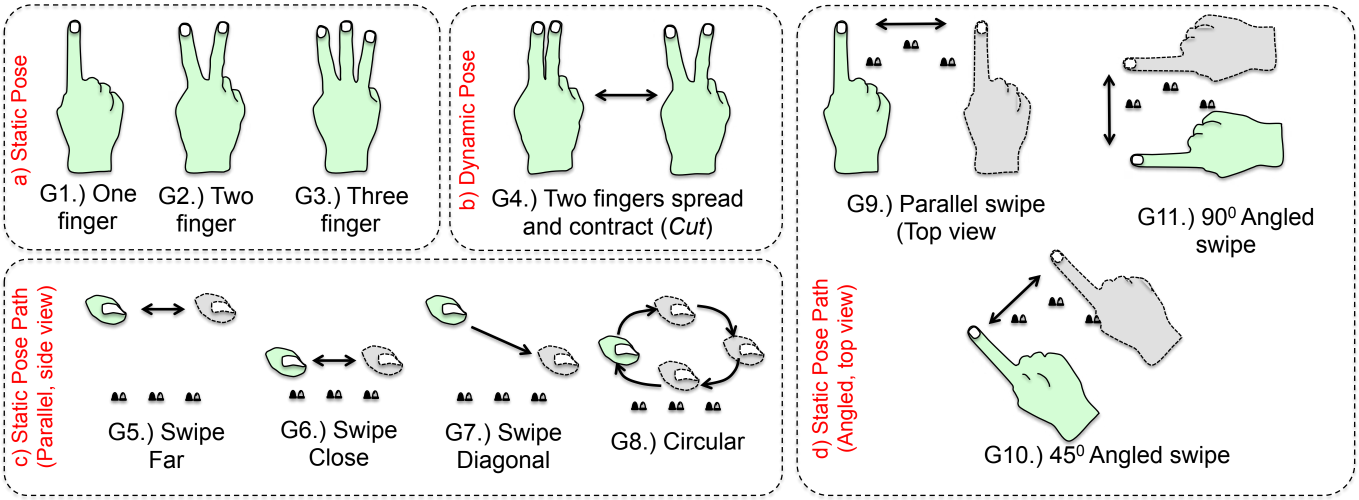


Figure 3: Set of gestures examined in the paper: a) Three *static poses*, b) One *dynamic pose*, c) four basic *static pose paths*, d) three angular variations in the direction of *static pose paths*.

Spatial Configuration		**Static Pose**	**Dynamic Pose**	**Pose Path**
3S3E	Linear	G1,G2,G3	G4	G5R,G6R,G7R G8CCW, G8CW
	Triangular	G1,G2,G3	G4	G7R, G10,G11
1S2E	Angular	G1,G2	G4	G5L,G7R

Table 1: Set of selected gestures and configurations.

Uniqueness: Gestures were selected to reflect unique features, such as poses, pose path, and dynamic poses. Only one iteration of mirrored gestures were selected (i.e. swipe left and swipe right are symmetric and only one of them is selected).

Relevance: Gestures with special relevance to a given configuration has been selected. For example, angled swipes (G10 and G11 of Figure 3) are only relevant to non linear arrangements, and only included as such.

Theory of Operation

We derived a generalised mathematical model to explain the principle of *zSense* operation. This model helps to make initial design choices in a *zSense* based gesture recognition system, by simulating it for various spacing and choice of components.

Received Intensity Model: Let us consider a *zSense* setup with m number of emitters, n number of sensors and a point target T at location $\vec{T_l}$. Figure 4a shows a simplified configuration of a reference emitter E_i $(i = 0, 1, 2, ..., m-1)$ at location $\vec{E_{li}}$ directed at $\vec{E_{di}}$ and sensor S_j $(j = 0, 1, 2, ..., n-1)$ at location $\vec{S_{lj}}$ directed at $\vec{S_{dj}}$. Let us take a ray received at sensor S_j, r, reflected from point target at $\vec{T_l}$, emitted from the emitter E_i. In order to calculate the received intensity at S_j, let's take optical characteristic-radiation pattern of emitter and sensor gain profiles to be, $I_i(\theta)$ and $G_j(\beta)$ respectively for the i^{th} emitter and j^{th} sensor.

The received intensity $f_{i,j}(\vec{T_l})$ at the sensor S_j from a ray emitted by emitter E_i with reference intensity W_i and reflected at target point T_l can be calculated using inverse-square law as,

$$f_{i,j}(\vec{T_l}) = \frac{I_i(\theta_{(i,\vec{T_l})})G_j(\beta_{(j,\vec{T_l})})W_i}{16\pi^2(|\vec{T_l} - \vec{E_{li}}|^2|\vec{S_{lj}} - \vec{T_l}|^2)} \qquad (1)$$

where
$\theta_{(i,\vec{T_l})} = \cos^{-1}\frac{(\vec{T_l}-\vec{E_{li}})\cdot\vec{E_{di}}}{|\vec{T_l}-\vec{E_{li}}|}$ and $\beta_{(j,\vec{T_l})} = \cos^{-1}\frac{(\vec{T_l}-\vec{S_{lj}})\cdot\vec{S_{dj}}}{|\vec{T_l}-\vec{S_{lj}}|}$

Note that $|\vec{E_{di}}| = |\vec{S_{dk}}| = 1$ since they are unit vectors and we assume the target to be radiating isotropically, and ignore the Lambert's cosine coefficient due to radial fall-off [8].

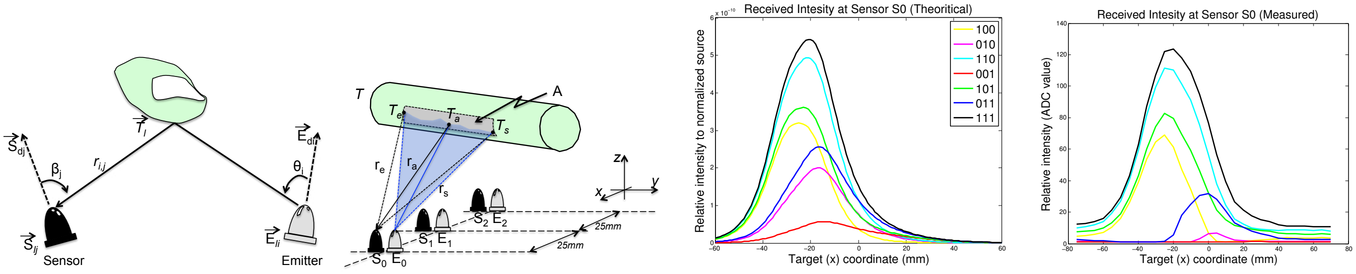
From equation 1, for each sensor-emitter combination, we can construct sensor-emitter cross intensity matrix with $m \times n$ dimensions, where each element $f_{i,j}(\vec{T_l})$ represents the power intensity received by j^{th} sensor due to the illumination by i^{th} emitter reflected by a target point at a given location $\vec{T_l}$.

Spatial light modulation pattern can be created by turning ON and OFF m emitters with p number of different patterns; where $p \leq 2^m - 1$. These patterns can be represented in a $p \times m$ matrix Λ, where each row represents a pattern and an element in a row, $\lambda_{i,j}$, represents whether the emitter is on or off using binary 1 or 0 in the given pattern.

Based on the pattern matrix Λ and sensor-emitter cross intensity matrix F, we can now compute the cumulative measured intensity matrix $A_{\vec{T_l}}$, with dimensions $p \times n$.

$$A_{\vec{T_l}} = \Lambda \times F \qquad (2)$$

Any member of $A_{\vec{T_l}}$, $\alpha_{k,j}(\vec{T_l})$, contains a measured intensity at j^{th} sensor by the cumulative sum of different ray compo-



(a) **(b)** **(c)** **(d)**

Figure 4: a) Configuration of a ray (r) emitted from emitter E, at location $\vec{E_l}$, reflected from a point T at location $\vec{T}$, received at a sensor S at location $\vec{S_l}$. Unit vectors $\vec{E_d}$ and $\vec{S_d}$ represent the directions of the emitter and sensor respectively. θ and β represents the existing and incidence angles of the ray r, b) Equally spaced linear configuration of three sensor-emitter units with IR rays reflecting from a cylindrical target, c) Profiles of the simulated values from one sensor, d) Measured values for the same sensor with same configuration and conditions.

nents emitted by different emitters, reflected from target point $\vec{T_l}$ according to the k^{th} SLM pattern.

$$\alpha_{k,j}(\vec{T_l}) = \sum_{i=0}^{i=m-1} \lambda_{k,i} f_{i,j}(\vec{T_l}) \qquad (3)$$

Usage of Theoretical Model: In order to demonstrate the theoretical model and its practical significance, we compared the experimental measurements with model-based simulations. For this purpose, we implemented 3 sensors and 3 emitters ($n = 3, m = 3$) configuration as shown in Figure 4b. We use off the shelf emitter, Optek technologies OP293[4], with emission angle of half power at $\pm 30^0$. Sensor used is Osram SFH203FA[5] half sensitivity angle $\pm 20^0$. We estimated the normalized $I_d(\theta)$ and $G_s(\beta)$ for emitter and sensor respectively to best fit the shape and parameters given in the data-sheet as follows,

$I_d(\theta) = \frac{1}{9\sqrt{2\pi}} e^{\frac{-\theta^2}{162}}$ and, $G_s(\beta) = \frac{1}{1+|\frac{\beta}{20}|^3}$

Due to the linear arrangement and the symmetry along $y = 0$ plane, we can assume each sensor-emitter units (i.e. S_0, E_0) are co-located along the line intersecting $y = 0$ and $z = 0$ planes. In the practical setup, each sensor unit was $25mm$ apart along the x axis.

We used a cylindrical rod with diameter $15mm$ as the target to imitate a human finger. In order to calculate the total reflected energy towards a given sensor i from the pattern k, we integrate over the complete reflective area of the target, A as shown in figure 4b. However, considering the symmetry of the target about $y = 0$ plane, the average reflected intensity will be representative of a point target T_a, lying on the $y = 0$. In order to easily compare the calculated and measured values at individual sensors, we kept the target constantly at $z = 50mm$ from the sensor plane ($y = 0$), and move it along the x axis in $5mm$ steps, taking 100 readings per sensor per pattern. Location of S_1E_1 is considered as the origin.

Three emitters E_i for $i = 0, 1, 2$ create a total of 7 patterns (i.e. $p = 7$). This is done by sequentially switching ON a single IR (001) to all three IR's (111). These patterns resulted in a measurement matrix A of 7 × 3, corresponding to 21 total measurements (i.e. 7 measurements from each of the 3 sensors) per any given $\vec{T_l}$ target point. As can be seen from Figure 4c and 4d, simulated values and measured values are correlated. Due to the space limitation, we only show the measured and simulated values of the received pattern for one sensor corresponding to each of the 7 emitted patterns. From the simulation, one can observe that once target displacement exceed 25mm (Figure 4c and 4d), curves get extremely converged. This span depends on the sensor directionality G_s and emitter relative radiant pattern I_d. Therefore, according to the choice of sensors and emitters, this model would be helpful to simulate the gesture sensitive area. In our prototype, we chose $25mm$ to be the ideal sensor-emitter unit displacement because we used the same sensors and emitters described above.

IMPLEMENTATION

zSense prototype is implemented using off the shelf IR sensors (SFH203FA) and emitters (OP293) and expected to work with maximum of 3 sensors and 3 emitter configurations. Sensor-emitter pairs are designed as single modules which can be plug into the main electronics unit.

Prototype

Figure 5a shows the block diagram of the *zSense*'s electronics unit. Sensors and emitters are driven using a commonly available Arduino Pro Mini[6] (16MHz, 5V) microprocessor to create the SLM pattern, which is modulated at 38kHz to reduce background noise. Figure 5b, 5c, 5d shows SLM patterns generated by linear 3S3E configuration. Patterns shown are (0,1,1), (1,0,0), (1,0,1) respectively. They were shot as the IR light reflected on a vertically standing sheet of white paper along the horizontal sensor line using an IR camera.

IR sensor input is amplified by a current to voltage converting amplifier (ST TL074 op-amp) followed by a phase lock amplifier (using TL074) to get a high signal to noise ratio (SNR). This significantly reduces the noise from external sources such as florescent lights. Our experiments were carried out directly under fluorescent lights, and the observed

[4] `http://goo.gl/d428fz`

[5] `http://goo.gl/sxTdLR`

[6] `http://arduino.cc/en/Main/ArduinoBoardProMini`

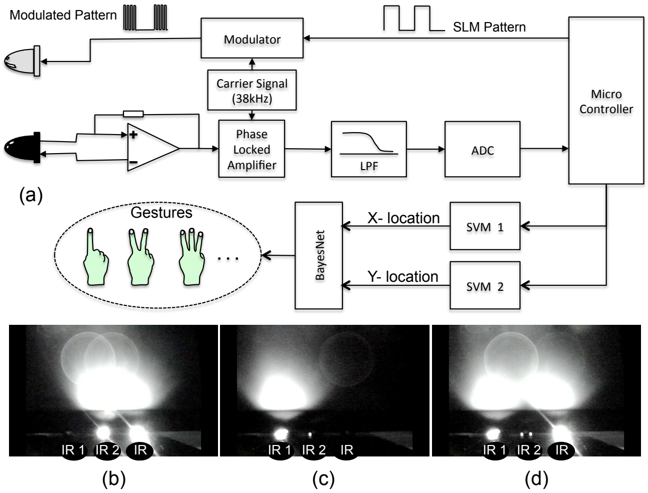


Figure 5: a) Prototype hardware block diagram, b),c),d) Three different SLM emitter patterns, (0,1,1), (1,0,0), (1,0,1)

effect was minimal. We have tested the system outdoors and the performance reduced only when the sensors were exposed to direct sunlight. We digitize the amplified signal using MCP3204 analog to digital converter (ADC) and transmit the data to the computer. ADC needs $10us$ conversion cycle, therefore each emitter pattern kept ON for $10us$ per sensor. Measurements from different patterns are combined into one data frame, and averaged over 10 frames. These frames were transmitted to the computer at rate of $50Hz$. This reduced acquisition rate contributes to low power consumption. Total emitter power consumed by *zSense* with 3S3E is $16.8mW$ and 1S2E $4.2mW$. Amplification and ADC stages consume about $3mW$ per sensor. Resulting total power for 3S3E to be $29mW$ and 1S2E to be $7.2mW$ (excluding microprocessor). Power saving mechanisms such as sleep modes can be introduced to save the power even further.

Software Classifier

A software classifier was implemented as a two stage process with two Support Vector Machines (SVMs) and a BayesNetwork algorithm provided by the Weka Toolkit [11]. First stage estimates the finger locations and second stage determines the exact gestures.

In the first stage, *zSense* module sends the measured IR light level of each sensor for each pattern as a 8 bit (ADC reading) numerical value. These readings are fed into the two Multi class SVMs which are trained to estimate vertical and horizontal level of the finger location respectively. Additionally, the number of fingers are determined by the same classifiers. For example, we trained two SVM classifiers using a rig to estimate 5 vertical (first classifier) and 5 horizontal levels (second classifier) for 1 finger, 2 vertical and 1 horizontal levels for two fingers and a single point for three fingers.

In the second stage, 20 consecutive finger locations are fed into a multinomial Naive Bayes classifier to determine the corresponding gesture.

EVALUATION

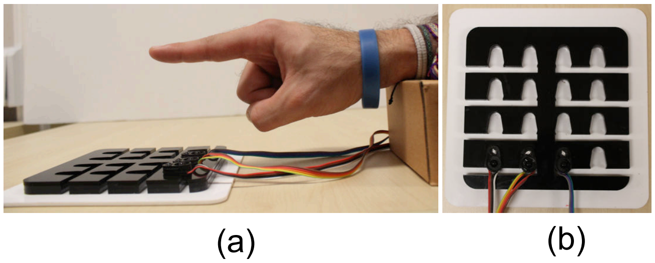


Figure 7: a) Side view of the study setup showing arm rest position and sensor grid, b) Top view of the sensor grid

The main objective of the experiment was to assess whether the current implementation of *zSense* is sufficiently accurate so as to ensure that future user studies with *zSense* enabled applications will yield unbiased data. Another goal of the study is to mix and match different gestures to different configurations, so that, in future, designers can use the findings of this paper as a key to create gesture interfaces using *zSense*. We describe the gesture set that was tested, the study method and the results.

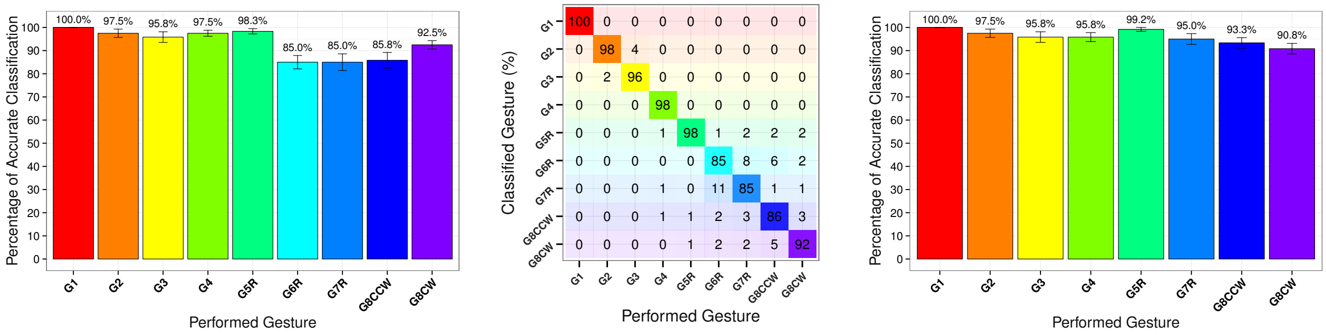
Method

Twelve (11 males and a female, age with min = 23, max = 36, mean = 28.3) participants took part in the study. All of them were right-handed and used their dominant hand to perform the gestures.

The experiment was conducted in three separate sessions, studying 3S3E linear, 3S3E triangular and 1S2E angular configurations respectively. For each session, the selected set of gestures were performed as shown in the table 1. In order to keep the relative sensor-emitter locations consistent between users and configurations, we used a grid assembly as shown in Figure 7.

Before starting each trial, the experimenter demonstrated the intended gesture to the subject. The subject was then given a few test trials to practice the gesture. Order of the sessions were counterbalanced and the gestures were randomized to eliminate bias. Subjects were instructed to only bring the finger to the sensitive region of *zSense* at the start of the gesture and to remove it once they finished. Since *zSense* can identify the presence of a finger, start and end point of the gesture were identified automatically. Participants were allowed to take a rest anytime during the study and average duration for the study was 40 minutes including the training and evaluation phases.

Training phase: During the training session, participants were briefed about the gestures they had to perform. In addition, they were instructed to adjust their gestures in order to capture variations of the same gesture. During the training session, we captured 20 instances of each gesture. For each gesture, finger location sequence was identified using two Support Vector Machine (SVM) models. Finger location training (i.e. training SVM modes) is not required for individual users. Therefore the two SVMs were trained by the experimenter before the user study.



(a) (b) (c)

Figure 6: Evaluation results for session 1 (Linear arrangement of *zSense* modules): (a) Classification accuracy for all 9 gestures in the session, (b) Confusion matrix between gestures, (c) classification accuracy after G6R is removed from the gesture set.

Evaluation phase: During the evaluation phase, participants were requested to perform 210 gestures. This corresponds to 21 gestures (9 in session 1, 7 in session 2 and 5 in session 3) and each gesture having 10 repetitions. We use this to calculate the accuracy of each gesture per each user with a BayesNet classifier.

Results

Accuracy Measures: The accuracy for each gesture is measured as a percentage of number of correctly classified gestures of that type. Our results showed that the accuracy of overall gesture recognition (all the gesture across all of configurations) is at 94.8% (SD=8.3%). In addition, mean percentage of confusion (false positives) are at 0.81% (SD=1.69%, max=10.8%). Results are discussed separately for each session.

Session 1: Linear Arrangement: Session 1 used three sensor and emitter units in a linear arrangement (Figure 2a) with $25mm$ space between them. We tested 9 gestures in this arrangement: one finger (G1), two fingers (G2), three fingers (G3), Cut (G4), Left to Right Swipe far (G5R), Left to Right Swipe close (G6R), Left to Right Swipe diagonal (G7R), Clockwise circle (G8CW), Counterclockwise circle (G8CCW). Short form of the name given within the brackets refer to the gesture number given in Figure 3 followed by letters to indicate the direction the gesture is performed. Figure 6a shows the average accuracy of classified gestures by 3S3E linear arrangement. All the gestures have accuracy over 85%, leading to overall accuracy result 93.05% (SD=9.61%).

Left to Right Swipe close (G6R, 85.0%, SD=10.0%) and Left to Right Swipe diagonal (G7R, 85.0%, SD=12.4%) had the least accuracy percentages. Using the confusion matrix shown in Figure 6b, it is evident that 10.8% of G6R is misclassified as G7R and 7.5% vice versa. One cause for the high confusion between close and diagonal right swipe (G6R and G7R) is most of the participants tends to follow a steep diagonal angle, so that the finger reaches close to the sensors in mid sweep. A solution to this problem would be either eliminate one of the gestures or be more specific in the instructions given to users.

At the post-hoc analysis, we removed the gesture G6R from the gesture data set and reclassified the data. Figure 6c shows the accuracy of the reduced gesture. Student's t-test reveals a significant improvement in G7R ($p < 0.01$) and G8CCW ($p < 0.05$). Average accuracy reduction in G8CCW was not statistically significant. Removing G6R can be justified because G6R is performed closer to the screen, where this might be confused with alternative technologies such as capacitive hover sensing.

Session 2: Triangular Arrangement: In session 2, three

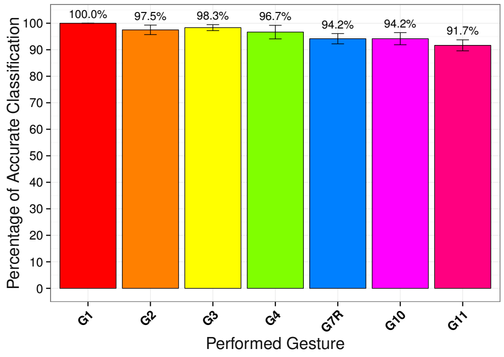


Figure 8: Classification accuracy of the gestures used in session 2 (Triangular arrangement of *zSense* modules)

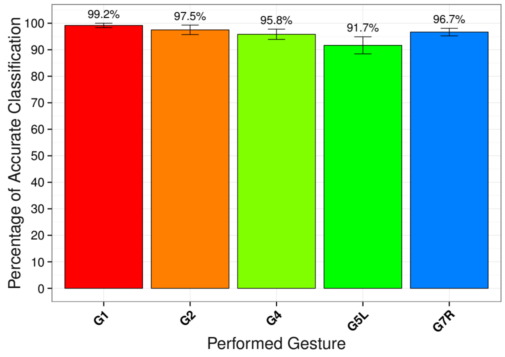
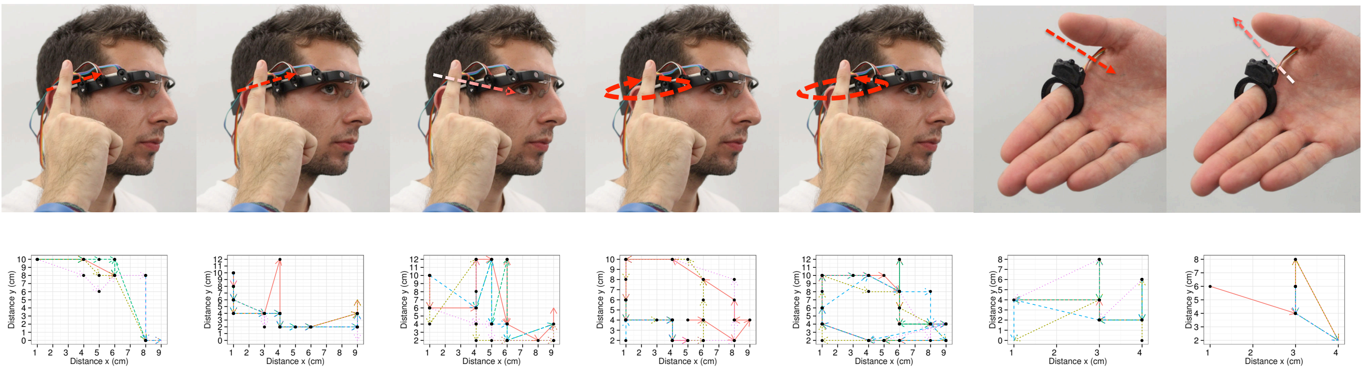


Figure 9: Classification accuracy of the gestures used in session 3 (Angular arrangement of *zSense* modules)



(a) G5R (b) G6R (c) G7R (d) G8CCW (e) G8CW (f) G5L (g) G7R

Figure 10: Detected $x - y$ finger locations of different static pose path gestures by *zSense* for five different instances

sensor-emitter units were arranged in a triangular configuration as shown in (Figure 2b right ($25mm$ spacing along the edges). We selected 7 gestures in this arrangement (Figure 3): one finger (G1), two fingers (G2), three fingers (G3), Cut (G4), Left to Right Swipe diagonal (G7R), 45^0Angled swipe (G10) and 90^0 Angled swipe(G11). Figure 8 shows the accuracies of each of the gestures in 3S3E triangular arrangement. All the gestures have accuracy over 91%, leading to overall accuracy of 96.0% (SD=6.77%). G11 (45^o angled gesture) has the least accuracy at 91.7% (SD=7.18%).

Session 3: Angular Arrangement: For the session 3, one sensor and two emitters (at an angle of 30^0) in a linear arrangement was used (Figure 2c1). In this session the following 5 gestures were analysed: one finger (G1), two fingers (G2), three fingers (G3), Cut (G4), Right to left swipe close(G5L) and Left to Right Swipe diagonal (G7R).

Gestures used in this session had an overall classification accuracy of 96.2% (SD=7.15%). Least accurate gesture was Left right swipe close (G5L) with mean percentage accuracy 91.7% (SD=11.15%). Results are shown in the figure 9.

DISCUSSION

We discuss the various aspects related to the concept, implementation and the evaluation of the *zSense*. We hope that these insights will help other researchers in the field to design spatially limited devices with greater input expressivity.

Implications of results: *zSense* is meant as a technology that provides greater input expressivity for devices with limited space, processing power and energy. We utilise non-linear spatial sampling to minimise the number of sensors and emitters required to identify shallow depth gestures. In this paper, we presented *zSense* configurations with a single sensor and two emitters capable of identifying five expressive gestures and two, three sensors - three emitter configurations capable of identifying eight and seven gestures with overall gesture recognition accuracy of 94.8%. From the evaluation results, we can conclude *zSense* is capable of identifying a significantly high amount of gestures with very few sensors and emitters, leading to high *gesture to sensor-density ratio*. Since the overall power consumption of NSS approach is lower, and the amount of data collected is smaller, *zSense* can be used as a durable and power efficient gesture recognition system in battery powered embedded systems.

Training Process: The current version of *zSense* requires two stages of training. First stage is trained per configuration and used to estimate finger location. This stage is required once for each configuration and does not depend on the individual user. Second stage is trained per user to extract gestures. However, the second level training is required by individual users. This is required to be performed only once at the beginning which also allows the user to familiarise with the new gestures.

Position Accuracy: *zSense* utilizes the non-linear spatial sampling in extracting a valid spatial representation of the finger location with minimal number of sensors. Sensitive region of *zSense* is very shallow with a resolution of about 20mm displacement between two adjacent detectable finger locations. A slight movement of fingers could lead to inaccurate estimation of a point. However, this does not affect *zSense*'s performance since multiple consecutive locations traced by a finger are able to counterbalance between accurate and erroneous location estimates and identify the correct gesture. This can be seen from Figure 10, where $x - y$ locations (of 5 instances) of different *static pose paths* are shown. These extracted $x - y$ locations were sufficient for a simple classifier to accurately identify the exact gestures. Gesture recognition accuracy can be improved even further by selecting only the required gestures as shown in Figure 6c.

Configuration-Gesture Relationship: *zSense* concept allows various configurations based on the number of sensors, emitters and their spatial displacement. Specific configuration can be chosen depending on the form factor of a particular device. For example, curved surface of a ring naturally provide the angular configuration required in 1S2E configuration of *zSense* (Figure 2c) and is able to recognize five expressive gestures with high accuracy. Similarly, 3S3E triangular configuration can be embedded on devices with flat squarish surfaces, such as a smartwatch (Figure 1a), enabling seven recognisable distinct gestures. The 3S3E linear configuration (Figure 1b), fits into devices with an elongated surfaces and able to distinguish eight gestures with high accuracy. Adding more sensors-emitters would not significantly contribute to the accuracy of gesture detection but would increase the interactive area.

zSense Application Space

Three potential application prototypes (Figure 11) are described here to demonstrate the broad application space of *zSense*. We hope these explorations will help peers in the field to inform future designs of expressive input capabilities on spatially limited devices.

***Extending interaction space of a Smartwatch*:** Integrating smartwatch with 3S3E triangular configuration can improve the interaction in two different ways. Firstly, it enables eyes-free interactions. For example, one can control a music application through swipe gestures without looking at the device, which is a useful feature when users attention is required for another task such as jogging. Secondly, gestures can add new possibilities to address the limitations of interactions. For example, playing a mobile game on a smartwatch could be difficult due to the limited space (fat-finger [28]), however, with *zSense*, one can use static poses and dynamic poses along with pose paths to trigger different aspects of the game (i.e. in a FPS game, G4 to shoot, G2 and G3 to change weapons). The smartwatch prototype is shown in Figure 11a, 11b.

***Enabling gesture interaction on a ring*:** We integrated the 1S2E configuration on the rim of a ring as shown in Figure 11c, 11d, 11e. The ring is able to sense five gestures, four of which can be easily performed with the thumb, enabling it as a private and personal gestural interface as proposed in [6]. Figure 11c shows a usage scenario of the ring interface as a slide controller. Additionally, Figure 11d shows how both hands can be used to interact using static poses (G1, G2) and dynamic pose (G4). Interacting directly with a glass can lead to fatigue over prolonged time. With *zSense*, arm can be kept in a relaxing position as shown in Figure 11e.

***Clip-on pin as a flexible and portable gesture interface*:** Figure 11f, 11g, 11h shows a clip-on pin prototype that uses the 3S3E linear configuration of *zSense*. This can be used as a flexible and portable mediating interface for other smart devices such as smartglasses, mobile phones, PDAs, etc. For example, it enables complex pose paths such as G8CW, G8CCW, that can allow fast scrolling a web page or a photo library through continuous cyclic gestures. It can can be attached to the pocket, collar and sleeves as shown in Figure 11f, 11g, 11h) to perform different gesture interactions according to the context.

LIMITATIONS OF THE CURRENT SYSTEM

Gestures identifiable through *zSense* depend on the spatial configuration and can be chosen based on the form factor of a particular device. For example, we observed that sensitive area is too narrow to perform G8 in both 3S3E triangular and 1S2E angular cases. Also, we observed that complex gestures such as G3 create significant ambiguities. As discussed under results, there is a high ambiguity between G5R and G6R.

In addition, sensing space of *zSense* is limited to 15cm from the sensors and only capable of accurately recognizing the finger gestures as opposed to high-resolution hand/finger/body gesture recognition provided by systems such as (Leap motion (LM), MS Kinect (MSK). However, LM/MSK uses focused sensing, which requires gestures to be performed from a certain distant from the sensor. This would be not desirable for applications that require subtle and private interactions. Since *zSense* uses non-focused sensors, it can detect very close proximity gestures.

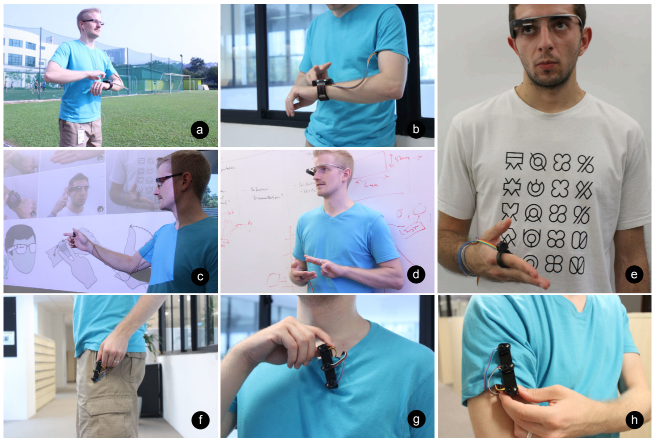


Figure 11: Application Space of *zSense*

zSense is unable to detect the precise location of user fingers with high accuracy. This is because, *zSense* was developed as a gesture sensing technology, which, detects a sparse *set* of finger locations to identify the gesture through a machine learning algorithm. Therefore, within its intended domain of gesture sensing, *zSense* works with high accuracy for gesture sensing applications. However, in our future steps, we intend to focus on the accuracy of the of the finger locations which would enhance and enrich the application scope of *zSense*.

Furthermore, since the current prototype of *zSense* uses IR sensors, limitations posed by IR technology persist in *zSense*. Current prototype utilizes noise resilient amplification techniques (i.e. phase lock amplification) to reduce the background noise to a minimum level. Even then *zSense*'s performance would be negatively affected by direct sunlight. However, through our experience with the system, we identified that this effect can be avoided by simply shading the sensors from direct sunlight. Additionally, the current version of *zSense* was developed using off-the-shelf components. However, by using surface mount components, customized printed circuit boards, etc., the system can be made much smaller and can be used in a more integrated manner.

CONCLUSION

In this paper we described *zSense*, a novel technology that effectively enables shallow depth gesture recognition on spatially limited devices such as smart wearables. We presented a non-linear spatial sampling (NSS) scheme which minimizes the sensor density, processing power and energy consumption required. The current implementation of *zSense* consumes less power ($1S2E < 8mW, 3S3E < 30mW$) and requires less processing capabilities (throughput rate $1S2E = 0.2kBps, 3S3E = 1.2kBps$). Our evaluations reported overall gesture recognition accuracy of 94.8% across three main configurations. This coupled with *zSense*'s ability to be used in different configurations, depending on the device form factor (as discussed), enhances its application scope with smart devices such as smartwatches, smartrings, smartglasses. Furthermore, the scope of *zSense* can be extended to mobile phones, tablets and other ubiquitous devices.

ACKNOWLEDGMENT

This work was supported by the International Design Center of the Singapore University of Technology and Design. Also, we thank all study participants for their time and valuable feedback.

REFERENCES

1. Ashbrook, D., Baudisch, P., and White, S. Nenya: Subtle and eyes-free mobile input with a magnetically-tracked finger ring. In *Proc. of CHI '11* (2011), 2043–2046.
2. Bailly, G., Mller, J., Rohs, M., Wigdor, D., and Kratz, S. ShoeSense: A new perspective on gestural interaction and wearable applications. In *Proc. of CHI '12* (2012), 1239–1248.
3. Baraniuk, R. G. Compressive sensing. *IEEE signal processing magazine 24*, 4 (2007), 118–121.
4. Bencina, R., Kaltenbrunner, M., and Jorda, S. Improved topological fiducial tracking in the reacTIVision system. In *IEEE CVPR Workshops* (2005), 99–99.
5. Butler, A., Izadi, S., and Hodges, S. Sidesight: multi- touch interaction around small devices. In *Proc. of UIST '08* (2008), 201–204.
6. Chan, L., Liang, R.-H., Tsai, M.-C., Cheng, K.-Y., Su, C.-H., Chen, M. Y., Cheng, W.-H., and Chen, B.-Y. FingerPad: Private and subtle interaction using fingertips. In *Proc. of UIST '13* (2013), 255–260.
7. Chen, K.-Y., Lyons, K., White, S., and Patel, S. uTrack: 3d input using two magnetic sensors. In *Proc. of UIST '13* (2013), 237–244.
8. Colao, A., Kirmani, A., Yang, H. S., Gong, N.-W., Schmandt, C., and Goyal, V. K. Mime: Compact, low-power 3d gesture sensing for interaction with head-mounted displays. In *Proc. of UIST '13* (2013), 227–236.
9. Duarte, M., Davenport, M., Takhar, D., Laska, J., Kelly, K., and Baraniuk, R. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine 25*, 2 (2008), 83–91.
10. Gustafson, S., Bierwirth, D., and Baudisch, P. Imaginary interfaces: Spatial interaction with empty hands and without visual feedback. In *Proc. of UIST '10* (2010), 3–12.
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter 11*, 1 (2009), 10.
12. Harrison, C., Benko, H., and Wilson, A. D. OmniTouch: Wearable multitouch interaction everywhere. In *Proc. of UIST '11* (2011), 441–450.
13. Harrison, C., and Hudson, S. E. Abracadabra: wireless, high-precision, and unpowered finger input for very small mobile devices. In *Proc. of UIST '09* (2009), 121.
14. Harrison, C., Schwarz, J., and Hudson, S. TapSense: enhancing finger interaction on touch surfaces. In *Proc. of UIST '11* (2011), 627–636.
15. Harrison, C., Tan, D., and Morris, D. Skinput: Appropriating the body as an input surface chris. In *Proc. of CHI '10* (2010), 453–462.
16. Hwang, S., Ahn, M., and Wohn, K.-y. MagGetz: customizable passive tangible controllers on and around conventional mobile devices. In *Proc. of UIST '13* (2013), 411–416.
17. Jones, B., Sodhi, R., Forsyth, D., Bailey, B., and Maciocci, G. Around device interaction for multiscale navigation. In *Proc. of MobileHCI '12* (2012), 83–92.
18. Ketabdar, H., Roshandel, M., and Yksel, K. A. Towards using embedded magnetic field sensor for around mobile device 3d interaction. In *Proc. of MobileHCI '10* (2010), 153–156.
19. Kim, D., Hilliges, O., Izadi, S., Butler, A. D., Chen, J., Oikonomidis, I., and Olivier, P. Digits:freehand 3d interactions anywhere using a wrist-worn gloveless sensor. In *Proc. of UIST'12* (2012), 167–176.
20. Kim, J., He, J., Lyons, K., and Starner, T. The gesture watch: A wireless contact-free gesture based wrist interface. In *Proc. of ISWC '07* (2007), 1–8.
21. MacKenzie, D. Compressed sensing makes every pixel count. *What's Happening in the Mathematical Sciences*, July (2009), 114–127.
22. Mitra, S., and Acharya, T. Gesture recognition: A survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 37*, 3 (May 2007), 311–324.
23. Nakatsuma, K., Shinoda, H., Makino, Y., Sato, K., and Maeno, T. Touch interface on back of the hand. In *proc. of SIGGRAPH '11* (2011), 39:1–39:1.
24. Nanayakkara, S., Shilkrot, R., Yeo, K. P., and Maes, P. EyeRing: A finger-worn input device for seamless interactions with our surroundings. In *Proc. of AH '13* (2013), 13–20.
25. Olberding, S., Yeo, K. P., Nanayakkara, S., and Steimle, J. AugmentedForearm: Exploring the design space of a display-enhanced forearm. In *Proc. of AH '13* (2013), 9–12.
26. Rekimoto, J., and Saitoh, M. Augmented surfaces: A spatially continuous work space for hybrid computing environments. In *Proc. of CHI '99* (1999), 378–385.
27. Ruiz, J., Li, Y., and Lank, E. User-defined motion gestures for mobile interaction. In *Proc. of CHI '11* (2011), 197.
28. Siek, K. A., Rogers, Y., and Connelly, K. H. Fat finger worries : How older and younger users physically interact with PDAs. In *Proc. of INTERACT '05*, Springer Berlin Heidelberg (2005), 267–280.
29. Wobbrock, J. O., Morris, M. R., and Wilson, A. D. User-defined gestures for surface computing. In *Proc. of CHI'09* (2009), 1083–1092.
30. Xiao, R., Laput, G., and Harrison, C. Expanding the input expressivity of smartwatches with mechanical pan, twist, tilt and click. In *Proc. of CHI '14* (2014), 193–196.